Analysing RNA-seq datasets to determine how pre-mRNA splicing is regulated by RNA binding proteins and cis-acting elements

Hugh-White, Rupert

Awarding institution:
King's College London

Analysing RNA-seq datasets to determine how pre-mRNA splicing is regulated by RNA binding proteins and cis-acting elements

Rupert Hugh-White

King's College London

PhD Supervisors: Reiner Schulz & Chad Swanson

A thesis submitted for the degree of

Doctor of Philosophy

May 2020

Declaration

Abstract

Most human genes undergo alternative pre-mRNA splicing to produce multiple transcript isoforms which often code for functionally distinct and tissue-specific products. Splicing factors interact with pre-mRNA and the core splicing machinery to control alternative splicing and regulate the generation of tissue-specific transcriptomes. For example, alternative splicing contributes to the function and homeostasis of the adaptive immune system. CD4+ T cells are an integral component of the adaptive immune system and regulate effector functions of immune cells towards diverse pathogens. Several splicing factors that contribute to CD4+ T cell function have been identified. However, the full splicing regulatory programmes characterising these and other immune cell types remain to be elucidated. Further, CD4+ T cells are the primary host target cell of HIV-1 infection. The HIV-1 lifecycle is regulated in large part through the host gene expression pathway. For instance, the HIV-1 RNA undergoes extensive alternative splicing mediated via the host splicing machinery. The study of processes such as these would benefit from development of improved methods for the inference of alternative splicing networks.

In this thesis, I have analysed RNA-seq datasets to understand how alternative splicing is regulated through the actions of RNA binding proteins and cis-acting RNA elements. Motif Activity Response Analysis (MARA) is an approach developed for the inference of tissue-specific regulatory transcription factors. I propose that MARA may also be effectively employed for the inference of regulatory splicing factors. To this end, I applied MARA for the novel use case of analysing splicing factors. I compared this Splicing-MARA (S-MARA) to a commonly used motif enrichment approach for predicting which splicing factors regulate a given splicing programme. For this purpose, I used a large-scale splicing factor knockdown data resource produced through the ENCODE project, in addition to a published CD4+ T cell activation timecourse. Despite its previous use, splicing factor motif enrichment analysis has not undergone a formal assessment. We found that this method has utility in identifying regulatory splicing factors, providing proof-of-concept for the use of motif-based methods in prediction of regulatory splicing factors. Counter to expectations, S-MARA had poorer performance in identifying regulatory splicing factors as compared to the motif enrichment method. As such, potential improvements to S-MARA are considered as future avenues for investigation.

Further, the RNA binding protein Sam68 was investigated using a knockdown approach to infer its genome-wide splicing targets during the CD4+ T cell activation process. This revealed a widespread role for Sam68 in regulating mRNA abundance, whilst only a limited number of genes showed Sam68-dependent alternative splicing. Finally, the regulation of the HIV-1 lifecycle by host RNA-binding proteins was investigated. We showed that suppression of CpG dinucleotides in the HIV-1 genome appears to maintain correct splicing of viral transcripts; whilst introduction of CpGs promotes use of a cryptic splice site which disrupts splicing, potentially mediated through the actions of host splicing factors.

Acknowledgements

This studentship was funded by Guy's and St Thomas' charity. Many thanks to my supervisors Reiner Schulz and Chad Swanson for their guidance throughout the duration of the project. My thesis committee, Rebecca Oakey, Sarah Teichmann, Eugene Makeyev, and Julie Fox provided useful advice to steer the direction of the project at multiple moments. Laura Hidalgo conducted the experimental work underpinning the analysis conducted in Chapter 5, and Irati Antzin-Andueza and Mattia Ficarelli conducted the experimental work described in Chapter 6.

Table of figures

List of tables

Abbreviations

APC = antigen presenting cell

BPS = branch point sequence

CLIP = cross-linking and immunoprecipitation sequencing

ENCODE = Encyclopaedia of DNA Elements

ESE = exonic splicing enhancer

ESS = exonic splicing silencer

FDR = False discovery rate

GO = gene ontology

hnRNP = heterogeneous ribinucleoprotein

LSV = local splicing variant

MAJIQ = Modeling Alternative Junction Inclusion Quantification

MARA = Motif activity response analysis

NMD = nonsense mediated decay

nt = nucleotide

PCA = principal component analysis

PSI = percent selection index

PSSM = position specific scoring matrix

PYT = polypyrimidine tract

RBNS = RNA-Bind-n-Seq

RBP = RNA-binding protein

S-MARA = Splicing Motif Activity Response Analysis

shRNA = short hairpin RNA

snRNA = small nuclear RNA

snRNP = small nuclear ribonucleoprotein

SRE = splicing regulatory element

TCR = T-cell receptor

WGCNA = Weighted Gene Co-expression Network Analysis

Chapter 1. Introduction

1.1 Splicing of pre-messenger RNA

Pre-messenger RNA (pre-mRNA) splicing is the process of intron removal coupled with ligation of the remaining exons, an essential process in the formation of mature mRNA (Sharp, 1994). Within eukaryotes, splicing is an essential component of the gene expression pathway, whilst within archaea it is mostly confined to tRNAs (Tocchini-Valentini et al., 2011), and in prokaryotes is considered rare and restricted to non-coding RNAs (Reinhold-Hurek and Shub, 1992).

1.1.1 Alternative pre-mRNA splicing

The use of alternative combinations of exons and splice sites from within a single gene contributes to the production of alternative mRNA and protein products (Sharp, 1994). Upward of 90% of human genes undergo alternative splicing (Pan et al., 2008; Wang et al., 2008), with ten or more isoforms often being generated from a given expressed gene (Djebali et al., 2012). In some instances, alternative splicing can generate large numbers of mRNA products from a single gene, such as with the *Drosophila Melanogaster* gene Dscam, from which tens of thousands of isoforms are generated (Schmucker et al., 2000). Complexity and prevalence of alternative splicing varies across species and shows greater divergence than patterns of gene expression (Barbosa-Morais et al., 2012). The prevalence of alternative splicing appears to correlate with organismal complexity, with a higher proportion of genes being alternately spliced in primates compared with other vertebrates (Barbosa-Morais et al., 2012), and in vertebrates relative to invertebrates (Kim et al., 2007). Alternative splicing is regulated in a highly tissue-dependent manner (Merkin et al., 2012) and contributes to the regulation of development and maturation processes (Kalsotra and Cooper, 2011). Mis-splicing of RNA is implicated in a variety of complex human disorders including cancer (David and Manley, 2010) and autism (Irimia et al., 2014); as well as single-gene disorders such as β^{+}-thalassaemia (Busslinger et al., 1981). Indeed, it has been postulated that disease associated mutations affecting splicing may be the most common form of hereditary mutation (López-Bigas et al., 2005).

A variety of alternative splicing patterns are recognised, with classically defined splicing categories including exon skipping, mutually exclusive exon usage, alternative 5' or 3' splice site usage, intron retention, and usage of an alternate last exon (Figure 1-1A) (Park et al., 2018). Usage of alternative first exons is an additional class of alternative splicing event (Figure 1-1A), but is driven through alternative transcriptional start site usage, rather than directly though the splicing machinery. More complex patterns of splicing also exist, with some examples shown in Figure 1-1B. Recently, Vaquero-Garcia *et al.* developed a novel approach for quantifying usage of alternative splicing variants of arbitrary complexity (Vaquero-Garcia et al., 2016), and found that ~37% of analysed genome-wide human alternative splicing events involved complex patterns of splicing not described by the more classical definitions.

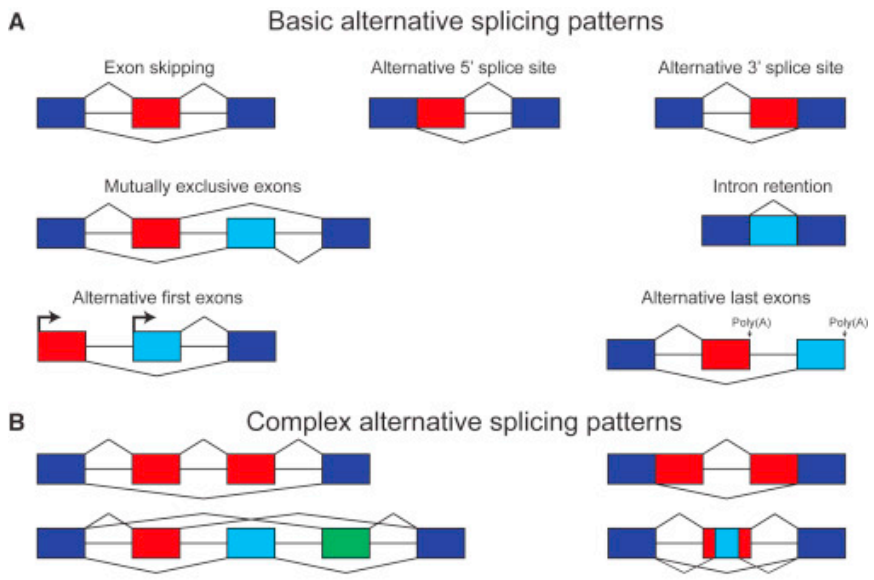

Figure 1-1. Categories of alternative splicing events. Dark blue boxes are constitutive exons, and red, light-blue, and green boxes are alternatively spliced exons. Poly(A) depicts alternative polyadenylation sites. Bold arrows show alternative transcription start sites. As published in (Park et al., 2018).

1.1.2 Alternative splicing in the generation of protein diversity

Alternative mRNA isoforms can code for proteins of divergent function and structure (Nilsen and Graveley, 2010). The functional consequences of alternative protein isoforms are wide-ranging, and include alterations to protein cellular localisation, ligand interactions, and enzymatic properties, amongst many others (Kelemen et al., 2013). However, the proportion of expressed alternative splicing events that are translated into distinct proteins genome-wide

is unclear. Several lines of evidence indicate a global correlation between alternative splicing and the generation of protein diversity. Weatheritt *et al.* employed ribosomal sequencing to study the ribosome-engaged fraction of the transcriptome (Weatheritt et al., 2016). With a focus on exon-skipping, this study identified a majority of mid-to-high abundance alternative splicing events as being detected via ribosomal sequencing, providing evidence that these splice variants may be translated. However, ribosomal occupancy does not guarantee a transcript will be translated, as the ribosome also functions in the quality control of mRNA (Inada, 2017). Liu *et al.* (Liu et al., 2017) utilised an alternative approach based upon correlating differential mRNA isoform abundance with differential protein abundance measured via a specialised mass spectrometry methodology (SWATH-MS), and found a high correlation between the two measures. To investigate potential effects of alternative splicing on protein-protein interactions, Yang *et al.* (Yang et al., 2016) used a yeast two-hybrid method to screen for protein interactions of alternative protein isoforms from several hundred human genes. This approach revealed that protein isoforms of a gene did not on average share more interaction partners than proteins produced from different genes, providing evidence that these alternative isoforms may contribute to diversity in the protein interactome.

Conversely, the role of error and stochastic noise as a major driver of transcript isoform generation has also been promoted as a hypothesis. Melamud *et al.* proposed a model based upon simulation of isoform expression in which production of a majority of isoforms are a consequence of stochastic noise (Melamud and Moult, 2009). They observed that the number of expressed isoforms per gene was a function of the number of introns and the level of expression, suggesting that alternative splicing may be driven stochastically during the transcription process. Consistent with this model, integration of multiple proteomic analyses indicates that for most highly expressed protein-coding genes, a single isoform is dominant across diverse tissues (Ezkurdia et al., 2015). Additionally, alternative exons show weaker evidence of selective pressure as compared to constitutive exons that form part of dominant isoforms (Liu and Lin, 2015; Tress et al., 2017a). Thus, whilst specific cases of alternative splicing leading to the generation of functionally distinct isoforms are well documented (Nilsen and Graveley, 2010), the proportion of alternative splicing events that contribute to the generation of proteomic diversity as a whole remains a contested topic (Blencowe, 2017; Tress et al., 2017b).

1.1.3 Alternative splicing in the control of gene expression

In addition to the generation of protein isoforms, alternative splicing contributes to the stability, transport, and translation of RNA molecules (Braunschweig et al., 2013). For instance, splicing is tightly linked with nonsense-mediated decay (NMD), a conserved quality control process that prevents aberrant expression of mutant or incompletely processed transcripts containing premature termination codons (PTC) (Hwang and Kim, 2013). NMD is enhanced by deposition of the exon junction complex (EJC), a large protein complex deposited 20-24nt upstream of exon-exon junctions during splicing (Hwang and Kim, 2013), and later removed by the ribosome during translation (Gehring et al., 2009). However, in cases of a PTC, frequently when present 50-55nt upstream of the terminal exon-exon junction (Nagy and Maquat, 1998), the EJC associates with upstream frameshift proteins to trigger termination of translation and release from the ribosome (Isken et al., 2008), with these transcripts subsequently being degraded through mechanisms involving exonucloeolytic or endonucleolytic cleavage (Schoenberg and Maquat, 2012). Alternative splicing is a source of PTC generation, and it has been estimated that 10-20% of such PTC-events represent cases of the directed control of steady-state transcript abundance initiated at the level of splicing (Pan et al., 2006). Interestingly, many of these splicing-NMD coupled events target RNA-binding proteins (RBPs) with splicing function, thus providing a regulatory feedback loop mechanism (Lareau et al., 2007; Plocik and Guthrie, 2012; Saltzman et al., 2008).

An additional layer of RNA quality control involves the retention of potentially aberrant transcripts within the nucleus. Export of mRNA from the nucleus is closely coupled with splicing, whereby recruitment of the transcription/export (TREX) complex occurs co-transcriptionally in a manner dependent on the splicing machinery (Masuda et al., 2005). The TREX complex facilitates transport of spliced RNAs through the nuclear pore via interactions with the NXF1 nuclear export receptor (Stutz et al., 2000). The presence of retained introns in incompletely spliced transcripts is known to prevent nuclear export, with such transcripts being retained in the nucleus where they are subsequently targeted for degradation in a mechanism that has not been fully elucidated (Yap and Makeyev, 2013). As with NMD, in some cases, such coupling of intron retention with RNA degradation is thought to provide a directed mechanism for controlling levels of gene expression. Yap *et al.* identified intron retention coupled to control of steady-state mRNA levels for a number of neuronal development related

genes, with this process facilitating the cell-type specific transcript expression (Yap et al., 2012).

1.1.4 Assembly and action of the spliceosome

Several broad mechanisms of splicing exist, including tRNA splicing (Randau and Söll, 2008), self-splicing (Pyle, 2016), and spliceosomal mediated splicing (Irimia and Roy, 2014), with the latter being the predominant mechanism by which most pre-mRNAs are spliced. The spliceosome is a large, nucelar, ribonucleoprotein complex (Will and Lührmann, 2011). Two spliceosomes of different composition exist, the U2-dependent spliceosome (or major spliceosome), and the U12-dependent spliceosome (or minor spliceosome) which is responsible for splicing of the rare U12 intron class (Patel and Steitz, 2003). The major spliceosome is composed of five small nuclear RNAs (snRNAs - U1, U2, U4, U5, and U6) bound to seven proteins to form a complex of small nuclear ribonucleoproteins (snRNPs) (Shi, 2017). Interaction between pre-mRNA and the spliceosome is mediated through binding of these snRNA and protein components with several core splicing sequence elements situated in the pre-mRNA (Figure 1-2). The 5’ and 3’ splice sites mark the donor and acceptor sites at which intronic sequences are cleaved and exonic sequences subsequently joined. The canonical splice site sequence is AG-GU, with the AG being prior to the cut site at the 3’ intronic end, and GU subsequent to the cut site at the 5’ intronic end (Figure 1-2) (Mount, 1982). Non-canonical splice site sequences are also used, although with relatively low frequency (Burset et al., 2000). The other constitutive splicing sequences are the branch point sequence (BPS), which is located upstream of the 3’ intron end and is less strongly conserved than the splice sites (Plaschka et al., 2019), and the polypyrimidine tract (PYT), which spans approximately 2-24nt downstream of the BPS, and is composed of pyrimidines with a bias towards uridine (Gao et al., 2008) (Figure 1-2).

Formation of the spliceosome into a conformation with catalytic activity is a dynamic stepwise process. The initial step is mediated via base-pairing of the U1 snRNP to the 5’ splice site, and of the 3’ splice site and PYT by the U2AF35 and U2AF65 components of the U2 auxiliary factor (U2AF) heterodimer, respectively (Figure 1-2) (Shi, 2017). Subsequently, the U2 snRNP base-pairs with the BPS in an ATP-dependent manner that is facilitated by U2AF at the 3’ splice site. This is followed by recruitment of the remaining subunits in the form of the U4/U6.U5 tri-snRNP. After further remodelling, involving removal of the U1 and U4 snRNPs, a final structure

with a catalytic active site is produced (Shi, 2017). This final spliceosome structure facilitates interaction between respective components situated at either end of the intron, and is thus said to be centred round the intron in a process of "intron-definition". However, in metazoans, the spliceosome initially forms around the exon through a process of "exon-definition", whereby binding of the U1 snRNP at a 5' splice site and U2 snRNP near to an upstream 3' splice site enhance one another in a cross-exon manner (Hertel, 2008). Introns are often many times longer than exons (Sakharkar et al., 2004), and initial assembly in an exon-definition manner over a relatively shorter distance is thus thought to facilitate the subsequent intron-focused spliceosome structure that must span a greater length of RNA (Moldón and Query, 2010).

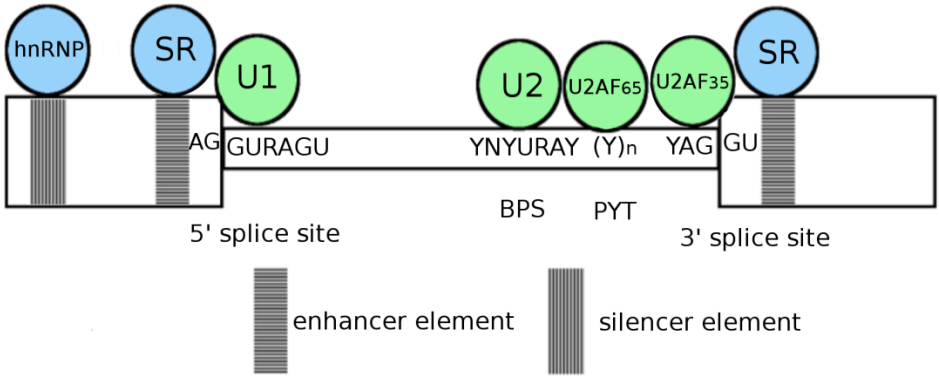

Figure 1-2. Selected components of the splicing reaction.

Larger boxes depict exons flanking an internal intron. Spliceosome components and core splicing factors are shown in green, auxiliary factors which contribute to alternative splicing are shown in blue. Consensus splicing elements are depicted: BPS = branch point sequence, PYT = polypyrimidine tract, Y = pYrimidine (C/U), R = puRine (A/G), N = aNy nucleotide. SR = Serine and arginine-rich protein, hnRNP = heterogeneous nuclear ribonucleoprotein. U2AF35/65 = components of the U2 auxiliary factor (U2AF) heterodimer.

After complete assembly, the multi-stage splicing reaction can be orchestrated by the spliceosome. The splicing reaction is centred round two phosphodiester transesterifications. The pre-mRNA is initially cleaved at the 5' splice site via transesterification of the BPS adenosine with the exonic splice site guanosine. This results in base pairing of these nucleotides and formation of a looped structure known as the intron lariat (Shi, 2017). The U2 and U4/U6 snRNPs are then involved in repositioning of the 5' site and BPS sequences (Shi, 2017). A subsequent transesterification reaction cleaves the intron at the 3' splice site, allowing covalent bonding between the 3' and 5' splice sites and resulting in joining of the

donor and acceptor exons. At this point, the intron lariat and associated snRNPs are released to complete the splicing reaction (Shi, 2017).

1.1.5 Auxiliary splicing factors and alternative splicing

The core splicing sequence elements show high levels of degeneracy in higher eukaryotes, and simulation studies suggest that they typically encode only half of the information required for accurate definition of exon/intron boundaries (Lim and Burge, 2001). Further, large introns often contain pairs of so-called 'decoy' splice sites, sequences with high similarity to the consensus splice sequences and which form the presence of pseudoexons (Krawczak et al., 1992). However, these sites are rarely spliced, highlighting the capacity for the splicing machinery to differentiate such false sites (Sun and Chasin, 2000). This capacity is thought to derive in part from additional sequence information present in the form of variable *cis*-regulatory motifs. These sequence motifs are present in both exons and introns and can function to either enhance or repress spliceosome assembly and splicing of a given exon (Figure 1-2). These features are referred to collectively as splicing regulatory elements (SREs), and can be sub-categorised as exonic splicing enhancers (ESE), exonic splicing silencers (ESS), intronic splicing enhancers (ISE), or intronic splicing silencers (ISS), depending upon their location and impact on splicing (Wang and Burge, 2008). Compared with the constitutive splicing elements, SREs have diverse sequence composition and higher levels of degeneracy (Wang and Burge, 2008).

SREs contribute to the regulation of splicing via facilitating the binding of *trans*-acting splicing factors through interactions with RNA binding domains (Dvinge, 2018). In turn, splicing factors modulate elements of splceosomal activity such as splice site recognition or assembly. In addition to constitutive splicing, *trans*-acting splicing factors also regulate alternative splicing, whereby the activity of a splicing factor may promote use of a given splice site pair at the expense of a neighbouring site. As such, splicing factors are a class of proteins which receive much attention in the study of alternative splicing mechanism. Two major classes of splicing factor are the serine and arginine-rich (SR) protein and heterogeneous nuclear ribonucleoprotein (hnRNP) families. The SR proteins are characterised by an arginine and serine rich C-terminal domain (RS domain), and an N-terminal RNA recognition motif (RRM) domain (Jeong, 2017). SR proteins show a preference for binding exonic purine-rich sequences (Änkö et al., 2012; Pandit et al., 2013; Sanford et al., 2004), and are generally considered to be

splicing enhancers which act through binding to ESEs (Jeong, 2017). hnRNP proteins are members of several different gene families but commonly contain variable combinations of four specific RNA-binding domains: the RRM, the quasi-RRM, a hnRNP K-homology (KH) domain, and an arginine/glycine rich (RGG) domain (Geuens et al., 2016). Members of the hnRNP family are classically recognised as splicing repressors which bind ESSs of diverse sequence composition (Wang and Burge, 2008). Whilst the classical roles of hnRNPs as splicing repressors and SR proteins as splicing enhancers holds true in many cases, knockdown studies have revealed large-scale induction of both exon skipping and inclusion directly mediated by SR proteins and hnRNPs (Fu and Ares, 2014; Llorian et al., 2010; Pandit et al., 2013). *Trans*-acting splicing factors can exert effects at numerous stages in the splicing reaction. SR proteins are known to interact with the U1 snRNP and U2AF to facilitate spliceosomal assembly (Black, 2003), and additionally may act to stabilize base-pairing of the BPS with the U2 snRNP (Shen and Green, 2006). Downstream steps of the splicing reaction are also subject to such regulation. For instance, hnRNP L has been shown to inhibit incorporation of the U4/U6.U5 tri-snRNP via binding to an ESS, thus acting to regulate spliceosome assembly rather than initial splice site recognition (House and Lynch, 2006).

The precise regulation of spliceosome activity is additionally controlled through the specific sequence context of SREs and the combinatorial actions of opposing or enhancing splicing factors. The actions of many splicing factors are dependent on their location of binding. For instance, hnRNP L has opposing enhancing or repressing activity when binding its CA-rich motif in exonic (Rothrock et al., 2005) or intronic (Hui et al., 2005) contexts respectively, a pattern that is also observed with other hnRNPs (Chen et al., 1999; Chou et al., 1999; Mauger et al., 2008). Similarly, SR proteins are known to have inhibitory effects when bound to upstream intronic regions through interfering with the downstream exon definition process (Erkelenz et al., 2013; Havlioglu et al., 2007).

Additionally, SREs exhibit combinatorial effects on splicing regulation. For instance, combinations of motifs are known to exhibit cooperative effects in promoting exon skipping (Han et al., 2005), and pairs of motifs are often found flanking exons in a fashion that synergistically promotes exon-definition (Ke and Chasin, 2010). Such combinatorial and context dependent effects of SREs have been found to follow common patterns and rules, leading to the concept of a “splicing code” through which alternative splicing may be predicted

as a function of *cis* acting motifs and the actions of *trans*-acting factors (Wang and Burge, 2008). However, whilst progress has been made towards modelling the splicing regulatory code via deep learning, this remains a challenging and open area of research (Barash and Vaquero-Garcia, 2014; Jha et al., 2017). Indeed, analysis of exonic sequences has led to the estimation that upwards of 1000 hexamers may have splicing regulatory function (Stadler et al., 2006). Thus, a given transcript will often contain many potential regulatory elements of which only a subset is utilised *in vivo*. An additional layer of information which may facilitate this precise utilisation and function of SREs is that of pre-mRNA secondary structures. Recent high-throughput *in vitro* analyses have identified many RBPs with strong preferences to specific RNA structural features (Burge et al., 2018). Further, a majority of transcripts are spliced co-transcriptionally (Brugiolo et al., 2013), providing opportunities for cross-regulation between the splicing and transcriptional machinery. Alternative splicing and transcription are kinetically coupled, whereby faster rates of transcription reduce the time for interaction between splicing factors and SREs, and therefore promote exon skipping, particularly at exons with weaker splicing signals (Kornblihtt et al., 2013). In turn, epigenetic factors such as exonic methylation levels can mediate the rate of transcriptional elongation via RNA polymerase II, and thus regulate levels of exon inclusion. Such an effect has been demonstrated with the transcription factor CCCTC binding factor (CTCF) (Shukla et al., 2011). Epigenetic modifications can also influence alternative splicing via the direct recruitment of splicing factors, as in the role of H3K46me3 in enhancing SRSF1 recruitment via the Psip1/Ledgf adapter protein (Pradeepa et al., 2012).

1.1.6 Control of splicing networks

The regulation of alternative splicing and its contribution to tissue specific transcriptomes is thought to be controlled in large part by the actions of key splicing factors (Jangi and Sharp, 2014). The concerted action of splicing factors is achieved through numerous mechanisms which give rise to a stable gene expression system whilst maintaining responsiveness to changing external stimuli (Jangi and Sharp, 2014). Negative autoregulation is a common feature of splicing networks, often mediated via coupling of splicing with NMD (Lareau et al., 2007; Ni et al., 2007; Saltzman et al., 2008). It can function to maintain steady sate expression levels by buffering against changes in transcription or protein stability (Nevozhay et al., 2009). Conversely, positive feedback commonly features in splicing networks and is associated with signal amplification and promotion of binary states such as cellular differentiation (Becskei et

al., 2001). For example, during neurogenesis, the transcriptional repressor REST is alternatively spliced by the splicing factor nSR100 to produce an isoform with reduced activity (Raj et al., 2011). This in turn increases nSR100 expression in acting as a positive reinforcement mechanism (Raj et al., 2011). Cross-regulation also appears to be a common motif in splicing regulatory networks (Fu and Ares, 2014). For instance, cross-regulation between several members of the hnRNP family is known to occur, and numerous hnRNP genes contain regulatory binding sites for other hnRNP proteins (Huelga et al., 2012).

Alternative splicing plays a key role in a number of signal transduction pathways (Fu and Ares, 2014). In particular, the activity of many SR proteins is dependent upon their phosphorylation status, which in turn can be regulated by extracellular cues such as growth factor signalling (Zhou et al., 2012) or stimulation of the T cell receptor (TCR) upon antigen presentation (Topp et al., 2008). Activation of splicing factors can result in signal amplification (Jangi and Sharp, 2014), whereby the downstream targets of splicing factor activation consist of both initial primary targets, in addition to secondary targets resulting from the altered splicing of those primary targets (Huelga et al., 2012; Jangi et al., 2014).

Further, alternative splicing frequently regulates the inclusion of disordered protein domains. Such disordered domains in turn modify the post-translational modification or protein-protein interactions of these alternatively spliced proteins (Buljan et al., 2013). As such, the regulation of disordered domains provides a mechanism through which alternative splicing can modulate protein-protein interaction networks. Indeed, disordered domains play a role in the phase separation of splicing factors in the formation of nuclear/splicing speckles (Itakura et al., 2018). Nuclear speckles are nuclear domains enriched in splicing factors and other RBPs and are located within interchromatin regions (Galganski et al., 2017). These speckles are dynamic structures, with exchange of components between speckles and the nucleoplasm or sites of active transcription. As such, nuclear speckles function in the assembly, modification, and temporary storage of pools of splicing factors, to facilitate dynamic splicing regulation.

1.2 Alternative splicing in CD4+ T cells

1.2.1 Role of CD4+ T cells in the adaptive immune system

The vertebrate immune system is composed of two complementary arms - the innate system and the adaptive system. The innate immune system comprises of physical defence barriers such as the skin, in addition to immune cells such as granulocytes (basophils, neutrophils, and mast cells), dendritic cells, and macrophages (Turvey and Broide, 2010). Innate immunity is mediated via recognition of conserved molecular patterns that identify potential pathogens and in turn activate defensive responses such as phagocytosis or activation of adaptive immune responses (Akira et al., 2006). Conversely, the adaptive immune system produces highly specific responses towards pathogens, and provides longer-lasting protection via the process of immune memory (Bonilla and Oettgen, 2010). The adaptive immune system is composed of B and T lymphocytes. B cell development takes place within the bone marrow, and mature B cells are responsible for mediating the antibody response via an antigen-specific immunoglobulin molecule - the B cell receptor (BCR) (LeBien and Tedder, 2008). T cell maturation occurs within the thymus, whereby thymocytes that originated from the bone marrow as haematopoietic precursors develop into two distinct lineages expressing a clonally restricted TCR – the TCRαβ+ and TCRγδ+ lineages (Schwarz and Bhandoola, 2006). The TCRαβ+ cells are the major lineage and are comprised of CD4+ T helper (T_h) cells and CD8+ cytotoxic T lymphocytes (CTLs) (Kreslavsky et al., 2010). CD4+ T cells are primarily associated with regulating the actions of other immune cells, including B cells and CD8+ T cells, whilst CD8+ T cells function to destroy virus-infected cells and tumour cells (Bonilla and Oettgen, 2010). T cell maturation is a complex multi-stage process which includes TCR gene rearrangement, selection via interaction with antigen-presenting cells (APC), and expression of lineage specifying surface molecules (e.g. CD4 or CD8) (Takahama, 2006). The specificity of the adaptive immune response is mediated by the extensive repertoire of expressed TCR and BCR variants, which is in turn generated through extensive somatic gene recombination (Roth, 2014). CD4+ T cells are implicated in various allergic and autoimmune conditions, and modulation of CD4+ T cells is therefore of therapeutic interest. For instance, dendritic cells can induce hyporesponsiveness towards specific antigen in CD4+ cells, which raises the potential for manipulation of hyporesponsiveness as a potential autoimmune therapeutic strategy (Maggi et al., 2015).

Some CD4+ T cells, such as regulatory T (T_{reg}) cells, are considered to be fully polarised effector T_h cells upon release from the thymus (Hsieh et al., 2012). A majority however, remain in a naïve state, and are capable of further polarisation into functionally distinct T_h subsets (Luckheeram et al., 2012).The process of a naïve CD4+ T cell polarising to a functional effector is driven by interaction of the TCR with its cognate antigen, which leads to so-called cellular "activation" and subsequent proliferation (Luckheeram et al., 2012). Antigen presentation to CD4+ cells is mediated by professional APCs, such as B cells, macrophages, or dendritic cells, which express a major histocompatibility complex (MHC) class II cell surface protein (Rock et al., 2016). The TCR engages with the antigen presented on the MHCII molecule, with CD4 acting as a co-receptor. For cellular activation to proceed, a co-stimulatory interaction between CD28 on the CD4+ T cell and a B7 protein on the APC is also necessary, as absence of co-stimulation leads to anergy and apoptosis as part of a mechanism to prevent inappropriate immune responses (Kalekar et al., 2016). TCR engagement causes phosphorylation of CD3, an intracellular component of the TCR, and of CD4, which initiates the T_h cell activation pathways through activation of Src family kinases by the intracellular component of the phosphatase CD45 (Courtney et al., 2018). This activation signal stimulates release of interleukin 2 (IL-2), a T cell growth factor, and upregulation of a subunit of the IL-2 receptor (CD25), promoting proliferation in both an autocrine and paracrine manner (Malek and Castro, 2010). After activation and the initiation of proliferation, CD4+ cells produce IL-4 and IFN-γ, in addition to IL-2, and are said to be in a T_{h0} cell state (Kim et al., 2001). An additional subset of activated CD4+ T cells that express neither IL-4 or IFN-γ, referred to as non-polarised activated T cells, are also recognised (Kim et al., 2001).

Helper T cells exhibit wide-ranging effects through the interaction with different immune cell types, and the specific function of a given T_h cell is determined through polarisation of T_{h0} cells towards distinct effector subsets. There are an increasing number of recognised T_h subsets, defined by their expression of key cytokines, chemokines, and associated receptor molecules, which in turn confer phenotypic effector functions (Saravia et al., 2019). Determination of T_h subtypes is determined through the combinatorial effects of manifold factors surrounding the T cell activation process including APC type, presence of specific costimulatory molecules, and composition of the local cytokine milieu (Saravia et al., 2019). In response to these stimuli, specification to each subset is then driven by the activations of master transcription factors

which orchestrate activation of subset-defining effector regulatory programmes (Tao et al., 1997).

The classically defined T_h subsets are T_{h1} and T_{h2} cells. Polarisation towards the T_{h1} class is driven through the combined actions of IFN-γ and IL-12 (Trinchieri et al., 2003). IL-12 is secreted by APCs in response to activation of pattern recognition receptors (Trinchieri and Sher, 2007). In addition to directly driving T_{h1} polarisation, IL-12 stimulates IFN-γ production by T cells and natural killer cells (NK) (Luckheeram et al., 2012), further driving T_{h1} polarisation. The T_{h1} master transcription factor T-bet is activated via STAT1 which in turn is activated by IFN-γ (Afkarian et al., 2002). T-bet both drives expression of T_{h1} promoting genes such as IL1-2 receptor β2 (IL2Rβ2), in addition to repressing the expression of genes involved in polarisation to other lineages, such as the T_{h2} promoting IL-4 (Djuretic et al., 2007). T_{h1} cells are associated with defence against intracellular pathogens through the pleiotropic actions of their major cytokines, IFN-γ, IL-12, and TNF-β, towards a variety of immune cells. For instance, IFN-γ stimulates macrophages to phagocytose intracellular pathogens, and IL-2 promotes the effector functions of CD8+ cells (Kim et al., 2006). Polarisation to the T_{h2} subset is controlled by IL-2 and IL-4, with these cytokines being secreted by T_{h2} cells in an autocrine positive feedback mechanism. The master transcription factor of T_{h2} cells is GATA3, which is upregulated upon the IL-4-dependent activation of STAT6 (Zhu et al., 2001). GATA3 acts to promote T_{h2} polarisation through various mechanisms including downregulating STAT4 to suppress T_h 1 differentiation (Usui et al., 2003), and upregulating the pro-T_{h2} transcriptional repressor Gfi-1 (Zhu et al., 2006). T_{h2} cells regulate the response to extracellular parasites and produce a variety of effector cytokines including IL-4, IL-5, IL-9, IL-10, IL-13, and IL-25, which collectively regulate the effector functions of numerous cell types such as basophils, eosinophils, mast cells, and B cells (Luckheeram et al., 2012). For instance, IL-4 stimulates the production of immunoglobulin E (IgE) by B cells, whilst IL-5 is involved in eosinophil activation, such as in the defence against helminths (Luckheeram et al., 2012). In addition to T_{h1} and T_{h2} cells, a number of other regulatory subsets exist. An increasing body of evidence indicates that T_h subsets exhibit a degree of plasticity, with intermediary phenotypes and reprogramming between different subsets, as opposed to the presence of terminally differentiated lineages (O'Shea and Paul, 2010; Zhou et al., 2009). This plasticity may allow flexibility in the immune response to different pathogens (Becattini et al., 2015), but inappropriate plasticity is associated with autoimmunity (DuPage and Bluestone, 2016).

1.2.2 Alternative splicing in CD4+ T cells

In addition to the actions of key transcription factors and transcriptional regulatory programmes, alternative splicing also plays a role in CD4+ T cell activation, polarisation, and homeostasis. From a genome-wide perspective, widespread alternative splicing has been documented in studies of activation using T cell lines (Ip et al., 2007; Martinez et al., 2012). Butte *et al.* demonstrated that stimulation of the TCR and CD28 in primary CD4+ T cells had synergistic effects on driving a programme of alternative splicing in several thousand genes, which was at least partially regulated via the upregulation of the splicing factor hnRNP LL (Butte et al., 2012). Ni *et al.* identified the widespread down-regulation of intron retention coupled with increased gene expression upon CD4+ T cell activation, particularly in genes in the proteasome pathway which is necessary for T cell proliferation and the release of cytokines (Ni et al., 2016).

One of the best studied examples of alternative splicing in CD4+ T cells is the transmembrane tyrosine phosphatase CD45 - encoded by the *PTPRC* gene. CD45 is expressed on the surface of all nucleated hematopoietic cells and has roles in proximal antigen recognition and cytokine-mediated signalling (Alexander, 2000). CD45 couples initial TCR engagement with downstream signalling by dephosphorylating tyrosine residues of the Src kinase Lck (Sieh et al., 1993). Exons 4, 5, and 6 of *PTPRC* are variable cassette exons encoding extracellular domains which undergo post-translational glycosylation (Alexander, 2000). Upon T cell activation, these cassette exons are preferentially skipped. This leads to a protein lacking the glycosylated domains, which in turn leads to CD45 homodimerization and reduced phosphatase activity, thereby decreasing TCR signalling in a negative feedback mechanism (Xu and Weiss, 2002). Expression of CD45 isoforms distinguishes different lymphocyte subsets, with naïve T cells expressing CD45 isoforms containing combinations of the three cassette exons, and activated cells expressing an isoform with all three exons spliced out (CD45RO) (McNeill et al., 2004). Differential expression of CD45RA (the isoform containing exon 4) and CD45RO is commonly used to differentiate between naïve and memory cells respectively, whilst expression of CD45RBC (which contains exons 4 and 6) is also seen in naïve cells (McNeill et al., 2004).

Control of CD45 alternative splicing is regulated by the combinatorial actions of a number of splicing factors (Figure 1-3) (Yabas et al., 2015). One of the first regulators of CD45 splicing identified was hnRNP L which binds to the activation response sequence (ARS), a motif present

in exons 4, 5, and 6, to repress splicing (Rothrock et al., 2005; Tong et al., 2005). Genome-wide analysis in a model T-cell-line has identified diverse targets under splicing control of hnRNP L (Cole et al., 2015). These effects were mediated by direct pre-mRNA binding of hnRNP L, in addition to indirect mechanisms involving epigenetic regulation (Cole et al., 2015). The hnRNP L paralogue, hnRNP LL, acts together with hnRNP L to repress splicing of CD45 exons 4 and 6 when its expression is upregulated upon T cell activation (Figure 1-3) (Oberdoerffer et al., 2008; Topp et al., 2008; Wu et al., 2008). Although hnRNP LL shows affinity towards the ARS, it neither binds to nor regulates splicing of exon 5 (Motta-Mena et al., 2010). Another splicing factor that contributes to skipping of CD45 alternative exons is PTB-associated splicing factor (PSF, gene name = *SFPQ*). PSF was first identified as a splicing factor through a screen of proteins which bind to the variable exons of CD45 (Melton et al., 2007; Topp et al., 2008). PSF binding is independent of hnRNP L and hnRNP LL and involves sequences outside of the ARS. The activity of PSF is regulated through phosphorylation by glycogen synthase kinase 3 in a manner sensitive to TCR signalling (Heyd and Lynch, 2010).

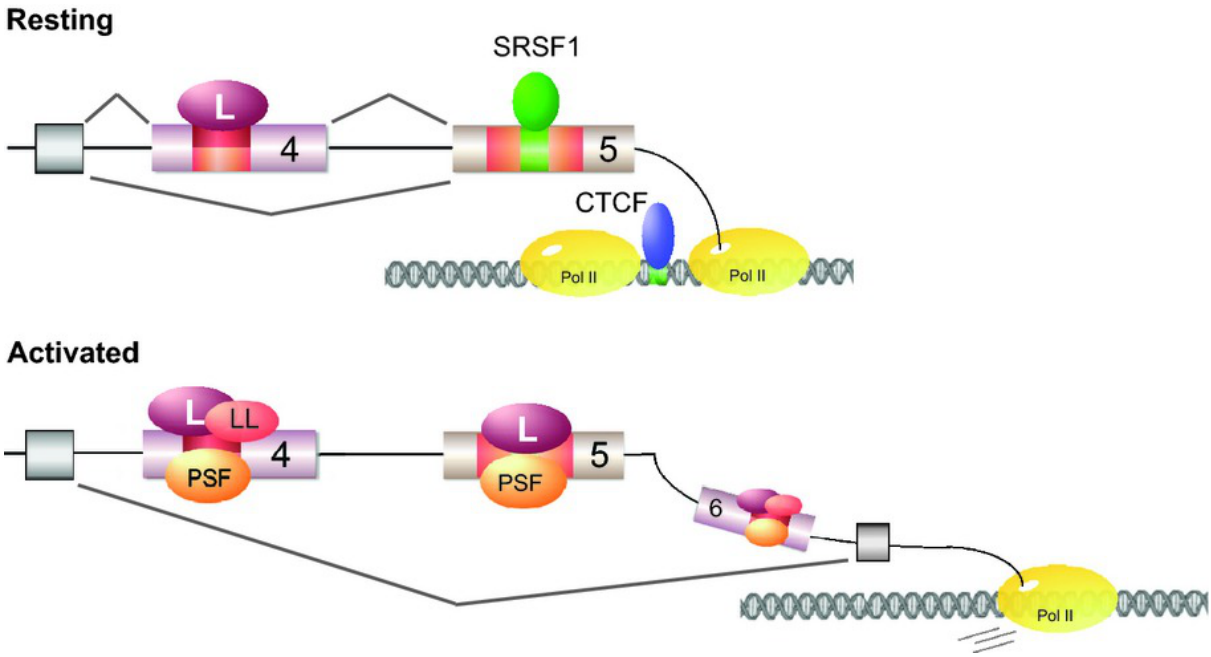

Figure 1-3. Control of alternative splicing at the CD45 locus. Transcription via Pol II at the CD45 locus is shown, with the speed of elongation being regulated via CTCF binding to exon 5. Co-transcriptional splicing of CD45 alternative exons is shown. The repressive action of hnRNP L (L), hnRNP LL (LL), and PSF, and enhancing action of SRSF1 and CTCF is depicted. The activation response sequence (ARS) is depicted in red and the ESE in green. As published in (Martinez and Lynch, 2013).

Several mechanisms of CD45 splicing enhancer activity have also been elucidated. The SR protein SRSF1 was identified as factor that promotes inclusion of exon 5 in resting naïve cells through binding to an ESE (Motta-Mena et al., 2010; Tong et al., 2005). SRSF1 and hnRNP L directly compete for binding to CD45 exon 5 to regulate mRNA inclusion levels. SRSF1 also regulates the alternative splicing of the CD3ζ chain of the TCR upon T cell activation, in a manner that enhances mRNA stability and translation through the regulated inclusion of a 3' terminal intron (Moulton and Tsokos, 2010). Inclusion of CD45 exon 5 is also regulated by the transcription factor CTCF. CTCF is recognised for its insulator activity (Singh et al., 2012), and also acts to decrease the rate of RNA polymerase II elongation by binding to exonic sequences, which in turn favours exon inclusion (Shukla et al., 2011). CTCF binding to CD45 exon 5 is regulated via its methylation status, with memory cells showing increased methylation at this locus which prevents CTCF binding and therefore enhances exon skipping (Shukla et al., 2011).

In addition to roles in mediating activity of proximal signalling components, alternative splicing contributes to a number of other processes in CD4+ T cell biology. For instance, TCR signalling strength was shown to modulate hnRNP A1 and hnRNP L splicing activity in a manner that regulated the ratios of polarisation towards T_{reg} or other T_h subtypes (Hawse et al., 2017). The T-cell-restricted intracellular antigen 1 (TIA1) is a regulator of FAS exon 6 skipping, and upon T-cell activation, it's activity is reduced in order to promote expression of an anti-apoptotic isoform (Izquierdo and Valcárcel, 2007; Liu et al., 1995). The regulation of apoptosis is critical for T cell homeostasis and development, such as in the control of central tolerance and the removal of autoreactive lymphocytes (Rathmell and Thompson, 2002). Alternative splicing is also important to the generation of functional cytokine receptors. For instance, IL-7Rα is alternatively spliced to produce a soluble and secreted form (Goodwin et al., 1990) necessary for survival of peripheral CD4+ T cells (Kondrack et al., 2003).

The precise function and regulatory mechanisms underlying the majority of alternative splicing events identified in CD4+ T cells have not been determined. A major goal in the study of immune function is to fully model and predict the splicing networks characterising processes such as T cell activation or polarisation (Martinez and Lynch, 2013). Understanding the details of splicing regulation in the immune system may have clinical implications. For instance, elucidating the mechanisms of immune disorders linked to mis-splicing, or predicting the

effects of polymorphisms on the splicing machinery (Cooper et al., 2009; Martinez and Lynch, 2013).

1.3 The HIV-1 lifecycle and its regulation by host RNA-binding proteins

1.3.1 HIV-1 is the etiological agent of the AIDS pandemic

Acquired immune deficiency syndrome (AIDS) is the clinical manifestation of late-stage human immunodeficiency virus (HIV) infection. The characteristics of AIDS, including increased susceptibility to common and opportunistic infection, are a result of a depletion of the CD4+ T cell population - the primary host target of HIV replication (Okoye and Picker, 2013). There are two main types of HIV - type 1 and 2, which originate from different cross-primate zoonotic events, with HIV type 1 (HIV-1) being the pandemic form (Sharp and Hahn, 2011). Antiretroviral therapy, where effectively distributed, has proven highly successful in reducing the number of deaths due to AIDS (World Health Organization; UNAIDS; UNICEF., 2011). Substantial challenges remain however, such as the large proportion of people predicted as being unaware of their HIV positive status (UNAIDS, 2014), and the continuing documentation of novel drug resistance mutations (Wensing et al., 2014). As such, HIV infection remains a major global health issue.

1.3.2 The HIV-1 lifecycle

The HIV-1 lifecycle is a complex multistage process (Figure 1-4). Entry to the host cell involves the fusion of the viral and cellular membrane via interactions between the HIV-1 envelope (Env) glycoproteins with host receptors CD4 (Klatzmann et al., 1984) and co-receptors CCR5 (Alkhatib et al., 1996) or CXCR4 (Feng et al., 1996). After entry into the cell, the HIV-1 RNA genome is reverse transcribed to cDNA by the viral reverse transcriptase (Hu and Hughes, 2012). After reverse transcription, the cDNA is incorporated into a nucleoprotein complex containing both host and viral proteins termed the preintegration complex (PIC) (Craigie and Bushman, 2012). The PIC is imported through the nuclear pore and the HIV-1 cDNA is integrated into the host DNA in a process catalysed by the viral integrase (Craigie and Bushman, 2012), with the host LEDGF/p75 protein acting as a cofactor (Ciuffi et al., 2005). The resulting integrated HIV-1 cDNA, referred to as the provirus, then utilises the host gene expression machinery to facilitate RNA transcription, 5’ capping, splicing, polyadenylation,

nuclear export, and translation (Karn and Stoltzfus, 2012). The viral Tat transactivator protein is required for efficient transcriptional elongation (Rice, 2017). Nuclear export of the incompletely spliced transcripts requires interaction between Rev, the Rev-response element (RRE) in the viral RNA, and host nuclear export factor CRM1 (Yi et al., 2002). The Gag polyprotein, which encodes the HIV-1 structural proteins capsid, matrix, and nucleocapsid, orchestrates the assembly of new virus-like particles at the plasma membrane (Bell and Lever, 2013). Budding of viral particles is mediated by the host ESCRT pathway (Usami et al., 2009), and is the process by which the viral lipid envelope is acquired. After budding, the final maturation process occurs which involves proteolysis of Gag and Gag-Pol polypeptides to yield the HIV-1 structural proteins and enzymes, and is catalyzed by the viral protease (Pettit et al., 1994).

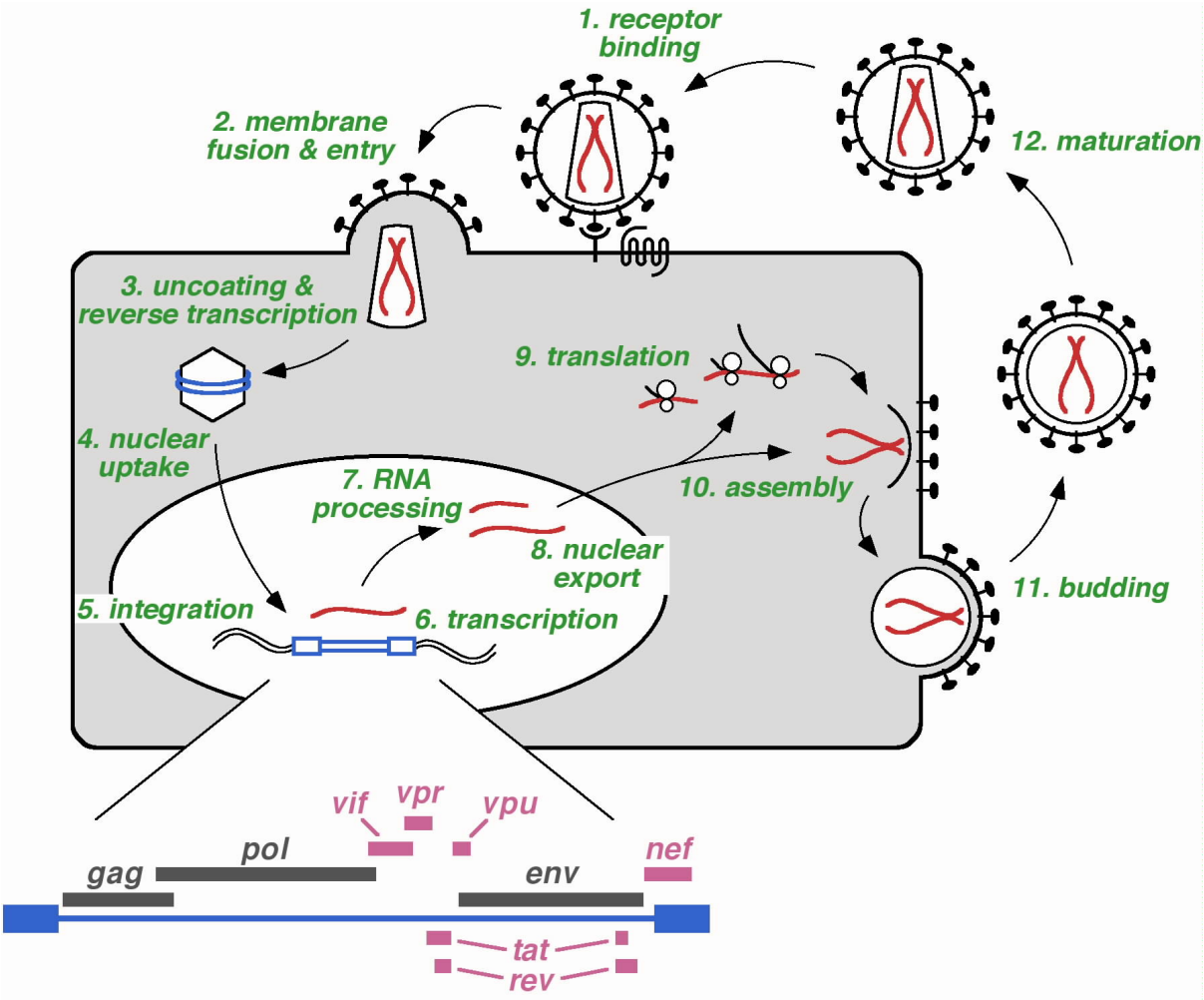

Figure 1-4. Stages of the HIV-1 lifecycle. 1) Glycoproteins present on the viral envelope bind with cell surface receptors. 2) The HIV-1 and host cell membranes fuse, allowing entry of the

viral core to the cytoplasm. 3) The viral capsid uncoats and the HIV-1 genome is reverse transcribed into cDNA. 4) The preintegration complex enters the nucleus via nuclear pore complexes. 5) The viral cDNA is integrated with the host genome forming a provirus. 6-9) The provirus utilises the host gene expression pathway for: 6) transcription, 7) RNA processing such as splicing, 5' capping, and polyadenylation, 8) nuclear export facilitated by the HIV-1 Rev protein, and 9) translation. 10) New virus-like particles are assembled, before 11) budding from the cell and 12) maturation. Schematic provided by Michael Malim.

1.3.3 HIV-1 depends upon the host splicing machinery

HIV-1 possesses a ~9kb genome containing nine open reading frames (ORFs) capable of producing 15 distinct proteins (Watts et al., 2009) (Figure 1-5). The integrated provirus is transcribed as a full-length pre-mRNA, under control of the HIV-1 promoter which resides within the 5' long terminal repeat (Karn and Stoltzfus, 2012). In addition to being the source of genomic RNA utilized for packaging into new viral particles, the full-length unspliced transcript acts as the pre-mRNA for subsequent splicing or is translated to produce the Gag and Gag-pol polypeptides after nuclear export (LeBlanc et al., 2013). Two classes of spliced transcripts are produced, the singly spliced, intron-containing transcripts, which are ~4kb in length and encode the viral envelope protein Env and auxiliary proteins Vif, Vpu, and Vpr, and the completely spliced transcripts encoding Tat, Rev, and Nef (Sertznig et al., 2018).

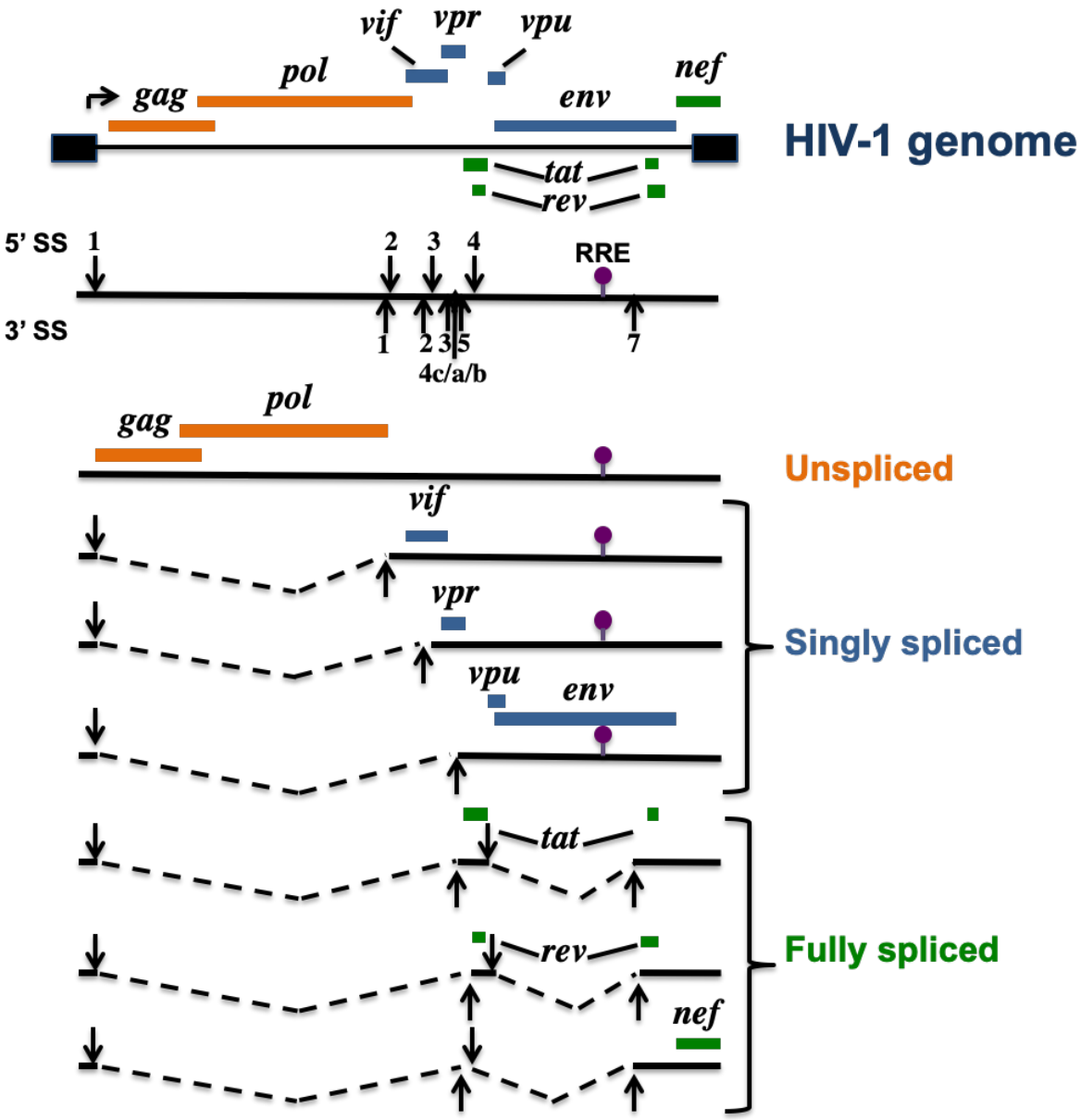

Figure 1-5. The HIV-1 genome, splice sites, and classes of viral transcripts. Top: The 9 HIV-1 ORFs frames are depicted, flanked by the 5' and 3' long terminal repeats (LTR) (black boxes). Initiation of transcription from the 5' LTR promoter is depicted. The canonical 5' and 3' splice sites (SS), in addition to the Rev-response element (RRE) are shown below. Bottom: The major HIV-1 transcript isoforms are depicted, grouped according to their degree of splicing. Figure from (Mahiet and Swanson, 2016).

To facilitate this extensive splicing, HIV-1 exploits the host splicing machinery, and indeed, screens for human HIV-1 dependency factors have identified numerous splicing factors as essential to viral replication (Sertznig et al., 2018). The HIV-1 genome contains numerous

canonical splice donor and acceptor sites, the coordinated use of which gives rise to the set of alternatively spliced viral transcripts (Figure 1-5). As with alternative splicing of host transcripts, *cis*-acting splicing regulatory elements play an important role in regulating the relative usage of donor and acceptor pairs through the actions of regulatory splicing factors.

Experimental approaches focused on introducing mutations to the HIV-1 genomic RNA have facilitated identification of splicing enhancer and suppressor elements for the majority of the canonical HIV-1 splice sites (Sertznig et al., 2018). For instance, when the A7 3' splice site is paired with D4, this results in removal of an intron which allows the formation of *tat, rev,* and *nef* mRNAs (Figure 1-5). The appropriate usage of A7 is predominantly controlled through two regulatory elements: exonic splicing enhancer three (ESE3) and exonic splicing suppressor three (ESS3) (Sertznig et al., 2018). ESE3 promotes binding of SRSF1 which stabilizes association of the U2AF65 subunit with A7 (Staffa and Cochrane, 1995). The ESS3 is a bipartite element comprised of two neighboring motifs (AGAUC and UUAG) and acts to repress A7 (Si et al., 1998). Suppressive activity is mediated through cooperative binding of hnRNP A/B proteins to the ESS3, which act to unwind local RNA secondary structures causing displacement of SRSF1 bound to ESE3 (Damgaard et al., 2002; Okunola and Krainer, 2009). Structural analysis of ESS3 has identified the presence of a stem loop structure which is critical for binding by hnRNP A1 (Rollins et al., 2014).

1.3.4 Human anti-HIV factors

Each step in the HIV-1 lifecycle relies upon numerous dependency factors, the result of viral adaptation in co-opting the host cellular environment. In addition, the host-pathogen coevolution process has resulted in numerous antiviral factors and reciprocal HIV-1 evasion mechanisms (Doyle et al., 2015). One class of human anti-HIV-1 proteins, termed restriction factors, have the potential to fully or partially restrict viral replication, but have mechanisms of evasion by wild-type HIV-1. HIV-1 restriction factors include TRIM5α, BST-2/Tetherin, members of the APOBEC3 protein family (APOBEC3G, APOBEC3F, APOBEC3D and some APOBEC3H variants) (Doyle et al., 2015), and the recently identified SERINC3 and SERINC5 (Rosa et al., 2015; Usami et al., 2015). These restriction factors, many of which are RBPs which directly bind to viral RNA, target different stages of the HIV-1 replication cycle, and all have known viral evasion mechanisms (Doyle et al., 2015). For example, BST-2, which encodes the Tetherin

protein, inhibits the cellular release of viral particles from Vpu-deficient HIV-1 by forming protein bridges between the viral and cellular membranes (Neil et al., 2008; Van Damme et al., 2008). In wild-type HIV-1 however, Vpu prevents trafficking of tetherin to the cell membrane and induces its ubiquitin-mediated degradation (Neil, 2013). In addition to these restriction factors, a further class of anti-HIV-1 proteins are those which reduce viral replication efficiency but do not appear to have direct viral evasion mechanisms. These proteins, referred to as HIV-1 resistance factors, are presumed to lack HIV-1 evasion mechanisms due to exerting a reduced selective pressure relative to the restriction factors (Doyle et al., 2015). The HIV-1 host resistance factors include MX2, IFITM1, IFITM2, and IFITM3, with SAMHD1 being a less clear case – having a demonstrated viral evasion mechanism in HIV-2 (Vpx) but being of unclear biological importance in HIV-1 infection (Doyle et al., 2015). As with the HIV-1 restriction factors, these resistance factors operate at different stages of the viral replication cycle. The IFITMs, for instance, interfere with membrane fusion to prevent viral cell entry (Compton et al., 2014; Lu et al., 2011; Tartour et al., 2014), whilst MX2 reduces nuclear entry and integration of viral cDNAs (Goujon et al., 2013).

1.3.5 Therapeutic targeting of host protein-HIV-1 interactions

As an essential component of the HIV-1 replication cycle, alternative splicing has been considered as a potential antiviral therapeutic target. Targeting of host dependency factors is appealing since this can increase the genetic barrier to viral drug resistance relative to the use of viral targets (Tang and Shafer, 2012). As an illustrative example, IDC16 was identified as a compound that interferes with SRSF1 activity, and was shown to reduce HIV-1 replication *in vitro* via disrupting viral splicing (Bakkour et al., 2007). Importantly, treatment with IDC16 did not affect the splicing of a panel of selected host genes.

In addition to targeting HIV-1 dependency factors such as the splicing machinery, directly facilitating the host innate immune response to the virus represents a further therapeutic strategy. The possibility of inhibiting the HIV-1 auxiliary proteins necessary for evasion of host restriction factors is an area of active research. For instance, small molecule inhibitor screens have been employed to successfully identify compounds which protect APOBEC3G from Vif-mediated degradation (Cen et al., 2010) such as through disrupting the formation of Vif-ubiquitin ligase complexes (Miyakawa et al., 2015). Similarly, a small molecule that enhances

Tetherin mediated restriction by inhibiting Vpu-Tetherin association and subsequent degradation has been described (Mi et al., 2015). Further elucidating the intricacies of host protein-HIV-1 interactions throughout the viral lifecycle is thus of interest in identifying new therapeutic avenues.

1.4 Profiling splicing networks and inference of regulatory splicing factors

1.4.1 Approaches to studying alternative splicing

Alternative splicing has been conventionally studied through reverse transcription polymerase chain reaction (RT-PCR) (Harvey and Cheng, 2016). The development of expressed sequence tag (ESTs) technology, a method for sequencing fragments of mRNA, allowed the identification of widespread alternative splicing across the genome (Modrek and Lee, 2002). In the mid-2000s, the development of splicing microarrays provided a further advancement and facilitated a high-throughput approach to the study of alternative splicing (Lee and Roy, 2004). However, microarray analysis is limited to the investigation of pre-defined splicing events. In 2008, several landmark studies employed short-read RNA sequencing (RNA-seq) to profile alternative splicing of both known and novel splicing events across the genome (Mortazavi et al., 2008; Pan et al., 2008; Wang et al., 2008). The high-throughput nature of RNA-seq and the ability to discover novel isoforms have led to it arguably becoming the gold standard for analysis of alternative splicing (Park et al., 2018). So-called “third generation” technologies, such as PacBio isoform sequencing, have allowed the sequencing of full length mRNA isoforms (Au et al., 2013; Sharon et al., 2013), however the lower-throughput nature of these techniques presents a challenge for quantification of relative isoform or splice event usage (Au et al., 2013).

Whilst RNA-seq is currently the gold-standard for profiling the contributions of alternative splicing in regulating the transcriptome, it does have limitations. For instance, RNA-seq captures steady-state transcript levels which are a function of both synthesis and degradation processes. As such, transiently expressed transcripts may be challenging to capture. Since splicing-NMD targets transcripts for degradation, the upregulation of such transcripts through alternative splicing can paradoxically decrease the total number of such transcripts through subsequent degradation. Sequencing of the nuclear RNA fractions provides a strategy through

which to more directly capture splicing processes such as intron retention and splicing-NMD (Zeng and Hamada, 2020). Further, native elongating transcript sequencing (NET-seq) facilitates sequencing of actively transcribed RNA, and thus allows direct investigation of co-transcriptional splicing (Nojima et al., 2015).

The popularity of short read RNA-seq has led to a multitude of algorithms for its use in the analysis of alternative splicing (Harvey and Cheng, 2016; Hooper, 2014). These approaches fall into several broad strategies, with one distinction being whether quantification is performed on the full transcript isoform or individual splicing event level. Quantification of full-length transcripts is challenging with short read technology, and sensitive to the use of reference transcriptome (Conesa et al., 2016), whilst the *de novo* reconstruction of isoforms is associated with challenges to both sensitivity and accuracy (Steijger et al., 2013). However, recent alignment-free quantifications allow rapid quantification of isoform abundances in addition to estimation of the associated uncertainty in assigning short reads to individual isoforms (Bray et al., 2016; Patro et al., 2017). Hybrid approaches that utilise initial isoform abundance estimations to infer relative usage of alternative splice events also exist, including the recently developed SUPPA2 (Trincado et al., 2018). Methods that aim to quantify individual splicing events, rather than full isoforms, may be further divided into those that focus upon the quantification of exons [e.g. DEXseq (Anders et al., 2012) or DDGseq (Wang et al., 2013)], splice junctions [e.g. MAJIQ (Vaquero-Garcia et al., 2016) or Leafcutter (Li et al., 2018)], or a combination of the two [e.g. SplAdder (Kahles et al., 2016), VAST-TOOLS (Tapial et al., 2017), or rMATS (Shen et al., 2014)]. Usage of junction-spanning reads is valuable in that such reads provide direct evidence for the usage of a particular splicing event. However, these reads represent only a reduced fraction of a total read library.

Commonly used metrics for quantification of alternative splicing are percent spliced in (PSI or Ψ), which quantifies the relative inclusion of a given exon or splicing event, and the selection or splicing index, which is a measure of the relative usage of a junction or exon when compared to the whole gene (Carazo et al., 2018). MAJIQ employs a percent selected index (also PSI) which captures the relative usage of each splice junction with all of its potential partner acceptor/donor junctions (Vaquero-Garcia et al., 2016). Many of these more recent methodologies demonstrate performance improvements over their predecessors [e.g. (Kahles et al., 2016; Shen et al., 2014)]. However, new approaches continue to be developed, and a

true gold standard approach cannot necessarily said to have been established (Carazo et al., 2018).

1.4.2 Profiling RNA-protein interactions

To understand the regulation of alternative splicing, the pre-mRNA targets of regulatory splicing factors must be determined. Methods for probing RNA-protein interaction may be RNA or protein-centric. A long-standing methodology is the gel electrophoresis mobility shift assay (EMSA), which involves combining mixtures of protein and nucleic acid (Ryder et al., 2008). Electrophoresis is used to separate these mixtures through a gel, whereby protein-RNA complexes will travel a different distance than un-complexed proteins or RNA, due to shifts in size, charge, and shape. Other methods involve use of UV cross-linking followed by protein immunoprecipitation, broadly termed cross-linking immunoprecipitation (CLIP). These methods facilitate a high-throughput a approach when combined with RNA-seq, such as RNA immunoprecipitation sequencing (RIP-seq) (Keene et al., 2006), and more recently, high-throughput cross-linking and immunoprecipitation sequencing (CLIP-seq) based methodologies (Hafner et al., 2010; Huppertz et al., 2014; Licatalosi et al., 2008; Van Nostrand et al., 2016). These approaches allow identification of *in vivo* protein-bound transcripts, of which a subset will represent functional interactions in the regulation of splicing or other RNA processing events. The identified RNA sequences are generally much longer than the RBP-bound region, which may only be several nucleotides (Carazo et al., 2018). However, bioinformatic (Zhang and Darnell, 2011) or methodological (Huppertz et al., 2014) adaptations to CLIP-seq allow individual nucleotide resolution of bound sites to be achieved. The resultant sets of protein-bound sequences can be used for determining RBP binding models, which in turn facilitate prediction of RBP binding potential in a given biological system without the use of further experimental work (Marchese et al., 2016). Alternatively, *in vitro* approaches based upon determining protein-RNA hybridisation efficiencies against pools of oligonucleotides can be used as input for the inference of such RBP binding models. These techniques include RNAcompete (Orenstein et al., 2016; Ray et al., 2013), RNA Bind-n-Seq (RBNS) (Burge et al., 2018), and RNA-SELEX (Jolma et al., 2010), which have all been used for high-throughput determination of both primary and secondary RNA structural preferences.

1.4.3 Motif models for RNA-binding proteins

RBPs show diverse RNA binding preferences, with multiple layers of sequence information determining affinity towards a given sequence. Reflecting this complexity, a wide range of approaches for the modelling of RBP binding preferences have been developed (Sasse et al., 2018). Classical methods for describing sequence motifs include the consensus sequence and the position specific scoring matrix (PSSM) (Stormo, 2000). PSSMs represent the relative frequencies of consecutive nucleotides in a linear sequence, with each position in the motif considered independently. The use of PSSMs can be extended to capture position interdependencies in the form of dinucleotides (diPSSM) (Riley et al., 2015), with more complex models capturing tri-nucleotide dependencies and higher-order interactions (Siebert and Söding, 2016). A number of implicit models based upon machine learning approaches have recently been developed which show high performance and are able to capture various higher-order complexities (Alipanahi et al., 2015; Budach and Marsico, 2018; Ghandi et al., 2014). Similarly, preferences towards binding RNA secondary structures are represented using models of varying complexity, from site-specific structural models describing a single structure per binding site, to models capturing higher-order position-specific structural preferences (Sasse et al., 2018).

1.4.4 Inference of regulatory splicing factors

1.4.4.1 Motif enrichment analysis

In order to fully understand the splicing network underlying a biological system, the key regulatory splicing factors must be determined. A common goal in the study of alternative splicing is therefore the prediction of which splicing factors and associated binding elements regulate alternative splicing in a given system. Various strategies that incorporate high-throughput splicing measurements with splicing factor binding data have been employed to this end (Carazo et al., 2018). A common strategy involves scanning RNA sequences surrounding alternatively spliced regions for the presence of potential regulatory elements, often through the use of PSSMs describing known splicing factor binding preferences. These binding predictions are then combined with an analysis of differential splicing of these regions under biological conditions of interest (Figure 1-6). Potential regulatory splicing factors can then be identified through analysis of over-representation or enrichment. Assessing for motif enrichment commonly involves comparing the distribution of predicted splicing factor binding

sites in differentially spliced regions relative to background genomic regions. Over-representation of motifs for a given splicing factor then provides evidence of a putative regulatory relationship (Figure 1-6).

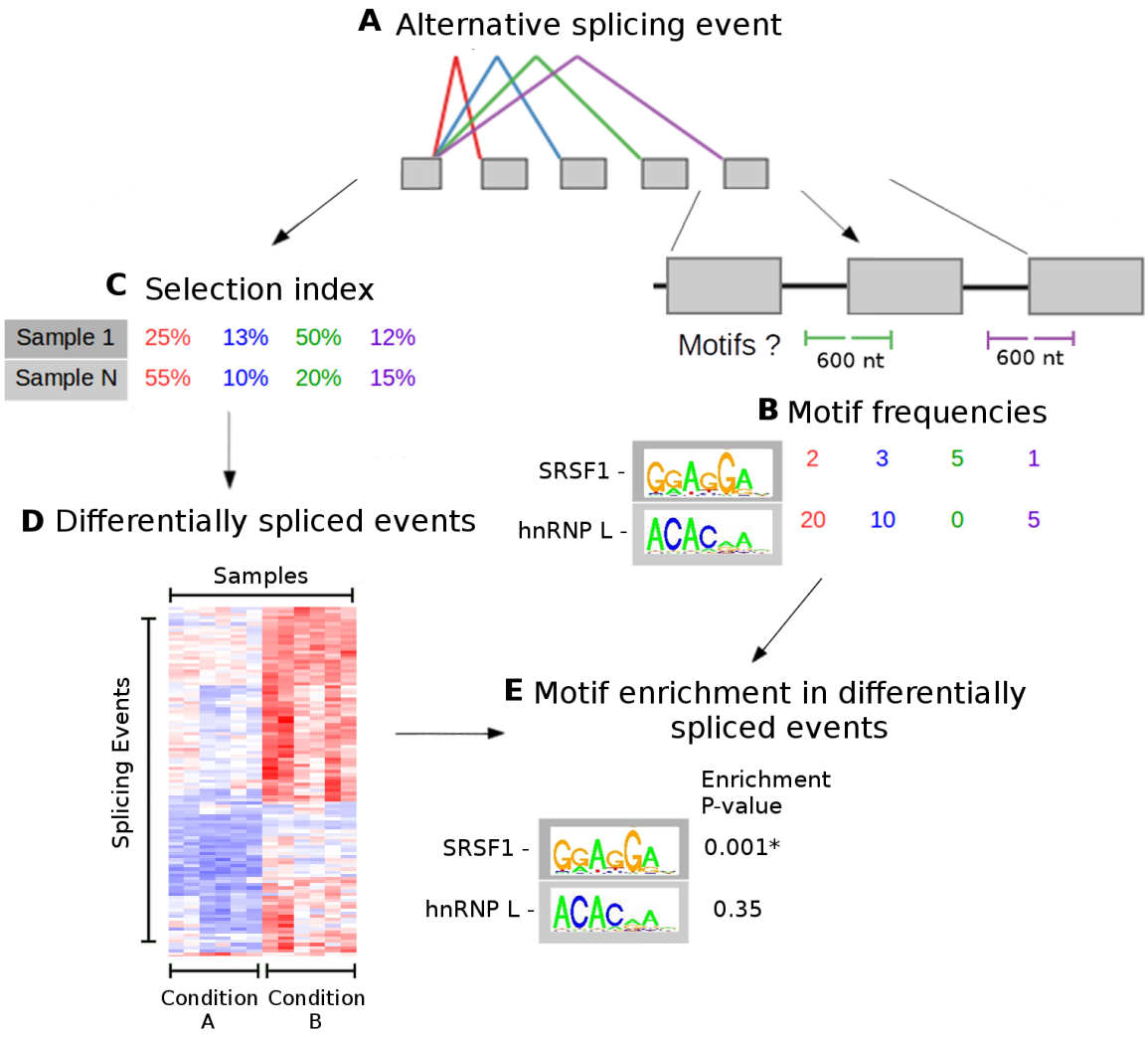

Figure 1-6. Schematic of a typical motif enrichment analysis workflow. (A) Alternative splicing events are defined genome-wide. **(B)** Regions surrounding alternative splice junctions are scanned for the presence of known or *de novo* identified motifs. **(C)** Usage of alternative splicing events is quantified per sample. **(D)** Differentially spliced events between biological conditions of interest are identified. **(E)** Enrichment for motifs within differentially spliced events is tested for using one of a variety of statistical approaches.

Such a motif enrichment workflow has been employed to study alternative splicing in diverse systems. Chen *et al.* studied alternative splicing during differential lineage commitment of hematopoietic stem cells (Chen et al., 2014). They employed a custom approach to identify

differentially spliced events, and scanned the corresponding exons and upstream and downstream introns separately for the presence of motifs representing binding preferences for 85 RBPs identified through a 2013 RNAcompete study conducted by Ray *et al*. (Ray et al., 2013). Enrichment for motifs in sequences flanking differentially spliced events was assessed via hypergeometric testing, and allowed identification of putative regulators of splicing during hematopoietic stem cell differentiation. Danan-Gotthold *et al.* identified a group of differentially spliced exons that were conserved across multiple human cancers and which were enriched for motifs of several splicing factors (Danan-Gotthold et al., 2015). Differentially spliced cassette exons were used for a *de novo* motif identification analysis before mapping the identified motifs to known splicing factor motifs and assessment for enrichment via hypergeometric testing. Validating this approach, knockdown of the identified splicing factors was found to result in a similar set of differentially regulated exons to those characterising the profiled cancers. Similarly, Sebestyén *et al*. found repeated enrichment for several splicing factor motifs amongst differentially regulated cassette exon regions in 11 human tumour types (Sebestyén et al., 2016). One of the identified splicing factors, MBNL1, was validated for a role in contributing to the cancer-related splicing landscape via extensive experimental follow-up. Sebestyén *et al*. used the motif scanning tool FIMO (Bailey and Elkan, 1994) – originally developed for analysis of DNA motifs, with the Ray *et al.* RNAcompete PSSM data (Ray et al., 2013) for motif identification, before motif enrichment testing was performed with a custom approach. Finally, Zhang *et al.* used RSAT (Nguyen et al., 2018) to perform both motif identification and enrichment analysis for the identification of regulatory splicing factors in MYCN driven neuroblastoma (Zhang et al., 2016). RSAT calculates expected motif occurrences using a background model which are then compared to the observed frequencies in sequences of interest to allow P-value estimation for motif over-representation (van Helden et al., 1998).

An alternative approach named CoSpliceNet was developed by Aghamirzaie *et al.* (Aghamirzaie et al., 2016). Their approach centred round identifying modules of co-expressed transcript isoforms and potential driver splicing factors, rather than directly measuring differential splicing. The MEME suite (Bailey et al., 2009) was then used for identification of *de novo* motifs separately in flanking upstream and downstream exonic and intronic regions of transcripts from each of the defined modules. The identified motifs were associated with splicing factors in a post-hoc fashion, which facilitated the identification of both known and candidate novel regulators of *Arabidopsis thaliana* embryonic development.

Other methods for the identification of splicing regulatory features have incorporated a comparative genomics approach (Sugnet et al., 2006; Voelker and Berglund, 2007; Yeo et al., 2007). Voelker and Berglund employed a comparison between seven eutherian mammals to identify conserved k-mers flanking splice junctions, a subset of which were found to be enriched in alternatively spliced regions (Voelker and Berglund, 2007). Again, a number of these conserved and putative splicing regulatory k-mers could then be mapped to known RBP motifs in a post-hoc manner.

1.4.4.2 Regression-based approaches

A further class of methods is focused on regression modelling of isoform or splicing quantifications as a function of sequence features (Das et al., 2007; Wen et al., 2013; Xin Wang et al., 2008; Zhang et al., 2012). These approaches tend to employ a *de novo* motif identification process in which a large sequence space is initially considered before potential functional sequences are defined which can then be mapped to experimentally defined motifs where possible. Xin Wang *et al.* used splicing microarrays to identify a set of differentially spliced cassette exons between human liver and heart as the basis for their approach (Xin Wang et al., 2008). Upstream and downstream exonic and intronic regions flanking each splice junction were considered as separate regulatory regions which were scanned for their frequency of all possible hexamers. Exon inclusion levels were then modelled as a linear function of the frequencies of each hexamer, with the functional effects of hexamers estimated as coefficients in the model. This model was run in an iterative process with a random selection of hexamers included each time, and performance of the model assessed via a least squares approach, minimising the difference between observed and predicted exon inclusion levels, and providing a method for hexamer scoring. The top 15 hexamers were considered as potential tissue-specific regulatory motifs, and these included known binding motifs for several splicing factors with previously identified roles in regulating alternative splicing in cardiomyocytes. Wen *et al.* implemented a similar regression-based approach, again focused on hexamers in putative regulatory regions flanking splice junctions of alternatively spliced exons across human tissues, but additionally considered potential combinatorial effects between pairs of splicing regulatory elements (SREs) (Wen et al., 2013). Due to the vast number of potential pairwise combinations of hexamers, a multi-stage model was used.

Initially, filtering of hexamers with low correlations to exon inclusion rates was performed, followed by further selection of potential SREs and SRE-pairs with an adaptive lasso regression approach. Significance of each SRE/SRE pair was then assessed using the ordinary least squares method. Using this workflow, the authors were able to predict sets of SREs and SRE-pairs that potentially contributed to tissue-specific alternative splicing. Again, the predicted SREs contained cases of known splicing factor motifs with previously recognised roles in regulating tissue-specific splicing, in addition to examples of predicted SRE pairs that have been previously demonstrated to exhibit combinatorial co-operative effects in regulating exon inclusion.

These regression-based methods have several potential strengths over motif enrichment-based procedures. Firstly, effects of different motifs are estimated as coefficients in a single model, and combinatorial effects may be explicitly modelled, rather than considering each motif in isolation as with an enrichment procedure. Further, since motif enrichment approaches rely solely on sequence data and require selection of a background sequence set for comparison, the results are sensitive to the choice of this background (Wen et al., 2013). Indeed, commonly applied corrections for underlying sequence composition such as GC content may introduce bias to the analysis by excluding AU or GC rich regulatory sequences (Wen et al., 2013). Further, since regression-based approaches directly incorporate quantitative data on splicing into the model, they are able to identify sequences which have a direct correlation with alternative splicing, and as such have been argued to be less sensitive to biases introduced by underlying sequence composition (Wen et al., 2013).

1.4.5 Developing methods for the inference of regulatory splicing factors

1.4.5.1 Motif Activity Response Analysis (MARA)

In 2009, the FANTOM Consortium published an approach for the integrative modelling of motifs termed Motif Activity Response Analysis (MARA) (The FANTOM Consortium et al., 2009). MARA is an approach for the inference of regulatory transcription factors using genome-wide gene expression quantifications and motif-based predictions of transcription factor binding sites in gene promoters. Since its original publication, MARA has been further developed and released as a web-based automated analysis tool (Balwierz et al., 2014). The MARA approach uses as input a matrix describing genome-wide counts of potential transcription factor binding sites within gene promoter regions. Promoter regions were

defined using information on transcription start sites acquired through deepCAGE sequencing data (Balwierz et al., 2009). Kilobase regions surrounding these transcription start sites are analysed for the presence and frequency of transcription factor binding sites using the MotEvo algorithm (Arnold et al., 2012) with 190 PSSMs that represent ~350 transcription factors. MotEvo considers sequence conservation levels across seven mammals combined with a model of constraints surrounding transcription factor binding site evolution. The result is a matrix with a score per promoter-motif pair which describes the sum of probabilities for all identified transcription factor binding sites within each promoter. This matrix is used as input for the MARA algorithm along with a corresponding matrix of promoter-matched gene expression quantifications. MARA employs a linear regression approach whereby the expression at each promoter is assumed to be a linear function of the binding site frequencies for all measured motifs/transcription factors. As with linear modelling approaches applied to splicing variation, the parameters that are estimated are coefficients that describe the activity of each motif in contributing to variation in the outcome variable. In this case, such "motif activities" describe the relationship between motif frequency at promoters and the average effect on gene expression across the genome, with the effects of different motifs being additive. Formalised, the equation that is solved is:

Equation 1. The MARA model.

$$Ep, s = \sum_m A_{s,m} \cdot N_{p,m}$$

where *Ep,s* is the sample and promoter specific expression, *As,m* is the sample and motif specific activity that is estimated, and *Np,m* is the number of motif counts per-motif at each promoter. This model is solved using singular value decomposition and an 80/20 cross-validation approach, whereby 80% of promoters are used as a training set and predictive performance on the remaining 20% is optimised whilst a Bayesian prior distribution is used to avoid overfitting. Ridge regression has also been used to solve this equation (Madsen et al., 2018). Both of these approaches are suited for the handling of sparse or collinear input data (Mandel, 1982; Zou and Hastie, 2005) - common features of motif count matrices owing to both the frequency of zero counts for transcription factor binding sites in some promoters, and the high correlations of count distributions for pairs of similar motifs. The fitted model typically explains a small but significant percentage of the variance in gene expression (Balwierz et al., 2014). However, the

true aim of MARA is in the estimation of motif activities, which represent the relative contribution of the motif-associated transcription factors to gene expression in a given biological sample. Changes in motif activity across samples from different biological conditions, such as from a time series experiment, provide evidence of potential regulatory roles in the control of gene expression (Figure 1-7).

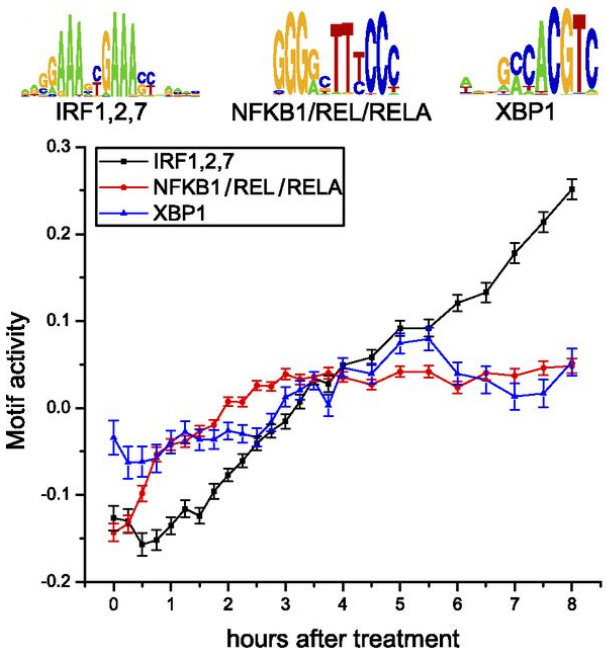

Figure 1-7. Example of MARA as applied to a time series of endothelial cell inflammatory response induced by tumour necrosis factor. Sequence logos depicting PSSMs for the three most significant motifs identified are shown. Two of the motifs are associated with binding of several potential transcription factors, as depicted. Error bars show standard deviations of inferred activities. The temporal profile of these motifs is suggestive of an active role of the associated transcription factors in driving TNF-induced gene expression changes. IRF1 and NF-κ
B are known regulators of the endothelial inflammatory response, validating the predicted

MARA has been applied to a wide-range of biological systems and has consistently been shown to recover known key regulators, in addition to predicting novel regulators (Balwierz et al., 2014). In its seminal application (The FANTOM Consortium et al., 2009), MARA was used to infer the identity, time-dependent activity profiles, and target genes for candidate transcription factor regulators of differentiation in the THP-1 cell line. A panel of 28 top

candidates were selected for experimental follow-up via small interfering RNA (siRNA) knockdown. A majority of the knockdown-induced modulations to gene expression were in line with the actions of these transcription factors as predicted by MARA, illustrating the value in the MARA approach. MARA has several strengths as a regression-based motif analysis strategy. Firstly, in estimating motif activity, MARA leverages the full, genome-wide data captured through an RNA-seq or microarray experiment, and does not require definition of a subset of differentially regulated events through thresholding of P-values. Additionally, MARA is centred on the estimation of motif activities per sample, facilitating the inference of relative activities of regulatory factors across biological conditions. The majority of regression-based methods applied for inference of regulatory SREs employ a *de novo* sequence analysis method. MARA, with a focus on the use of experimentally defined PSSMs associated with regulatory factors *ab initio*, identifies motifs of interest that are more directly interpretable. In contrast, *de novo* SRE/motif approaches lead to the generation of large numbers of candidate motifs of unknown function, some of which may influence splicing through indirect mechanisms such as through modulating the local sequence context surrounding splicing factor binding sites or by influencing RNA secondary structures. Finally, MARA is flexible with regards to the length of input motif, and does not require selection of a k-mer length in order to narrow the number of searched sequences to a tractable space.

1.4.5.2 Proposing a novel analysis approach – Splicing Motif Activity Response Analysis

To date, MARA has been applied to model the regulation of gene expression by transcription factors and microRNAs, in addition to the regulatory roles of transcription factors in influencing chromatin state (Balwierz et al., 2014) or enhancer occupancy (Madsen et al., 2018). However, MARA has not been leveraged for the study of splicing thus far. Whilst various mechanistic differences exist between the control of alternative splicing and of transcription, there is a commonality between the processes that should render splicing amenable to investigation with MARA. As variation in gene expression is in part driven by the presence of transcription factor binding sites within gene promoters, variation in splicing is in part driven by the occurrence of splicing factor binding sites flanking alternative splice junctions, and by the concerted tissue or condition specific actions of these splicing factors.

1.5 Thesis aims

Alternative splicing is both widespread throughout the genome and highly regulated across tissue-types and within biological processes from development to disease. For instance, the splicing of several loci critical to proper development and function of CD4+ T cells, which are key regulators of the cellular adaptive immune response, is well described (Yabas et al., 2015). Genome-wide patterns of differential splicing in CD4+ T cells, such as upon stimulation of the TCR, are consistently observed in *in vitro* studies (Ip et al., 2007; Martinez et al., 2012). The elucidation of how these networks of gene splicing are controlled through the actions of splicing factors remains an important goal, and will ultimately improve understanding of how T helper cells act to influence the adaptive immune response. CD4+ T cells are also the primary host cell of the HIV-1 retrovirus. The lifecycle of HIV-1 is dependent upon the interactions between viral components with numerous host proteins. Generation of new viral particles relies upon the host gene expression pathway, and thus involves the actions of RBPs in regulating processes such as the splicing, nuclear export, and translation of HIV-1 transcripts (Karn and Stoltzfus, 2012). Therefore, splicing factors and other RBPs are common HIV-1 dependency factors. The study of such HIV-1 dependency factors, in addition to host restriction factors with HIV-1 repressive properties, represents a potential avenue for development of new therapeutic strategies.

RNA-seq has become one of the most powerful approaches to studying alternative and differential splicing due to its transcriptome-wide and high-throughput properties, and the capacity to uncover novel splicing variants. In order to fully characterise the programmes of differential splicing identified through RNA-seq studies, knowledge of the mechanisms of spliceosomal control are needed. To this end, a common analysis strategy is the integration of RNA-seq-derived differential splicing profiles with models of RBP binding preferences, commonly in the form of PSSMs (Carazo et al., 2018). This strategy allows inference of putative regulatory splicing factors in a given biological system. In the analogous field of transcription factor biology, such strategies are arguably more sophisticated, whereby quantitative models of transcriptional activity and transcription factor binding site predictions are commonly applied to infer key regulatory interactions (Balwierz et al., 2014; The FANTOM Consortium et al., 2009). Improved methods for the inference of regulatory splicing factors could be applied to any field of biology in which the mechanisms of differential splicing are of interest.

Here, I propose to implement a workflow for the application of MARA to model splicing factor motif activity (S-MARA) and to assess the viability of this approach for the inference of regulatory splicing factors within a given biological system. The power of MARA in predicting regulatory transcription factors has led to its widespread use, and the strengths of the approach warrant an investigation into its application to the study of splicing. The performance of S-MARA will be assessed using a large-scale knockdown project resource generated through the ENCODE project (Burge et al., 2018; Davis et al., 2018; Nostrand et al., 2018). S-MARA will be compared to a motif enrichment analysis approach - which has been a frequently employed method in recent years for inference of alternative splicing regulators in a given biological system (Carazo et al., 2018). Additionally, to investigate the value of S-MARA in uncovering both known and promising novel candidate regulatory splicing factors, I will apply the approach to a timecourse of CD4+ T cell activation and polarisation. As discussed, the CD4+ T cell activation process is characterised by widespread regulation of alternative splicing. However, the function and regulatory factors controlling this programme of alternative splicing are only partially understood, and the further elucidation of this splicing network is an ongoing research aim.

In this thesis, I aim to analyse RNA-seq datasets to understand how alternative splicing is regulated through the actions of RNA binding proteins and cis-acting RNA elements. I propose that Motif Activity Response Analysis may be effectively applied for the inference of regulatory splicing factors. I therefore aim to apply this, as well as additional methodologies, to study the actions of RNA-binding proteins in the context of CD4+ T cell activation and the HIV-1 lifecycle.

The specific aims of this thesis are:

1. Implement an analysis workflow to perform Splicing Motif Activity Response Analysis (S-MARA).
2. Benchmark the performance of S-MARA in inferring regulatory splicing factors using splicing factor knockdown-RNA-seq data.
3. Further understanding of the roles of key splicing factors in regulating alternative splicing in CD4+ T cells:
   3.1. Apply S-MARA to predict known and novel candidate regulators of splicing during the CD4+ T cell activation process.

3.2. Profile the role of the RNA-binding protein Sam68 in regulating alternative splicing in CD4+ T cell activation.

4. Investigate the effects of CpG dinucleotides on the splicing of HIV-1 transcripts.

Chapter 2. Materials & Methods

2.1 RNA-seq pre-processing

Initial quality control of RNA-seq FASTQ files was performed using FastQC v0.11.5 (Simon, 2010), followed by compilation of individual quality reports with MultiQC v1.4 (Ewels et al., 2016). The Trim Galore! wrapper to Cutadapt v0.4.2 (Martin, 2011) was used for removal of adapter sequences and low quality (phred score < 20) read ends, with resulting reads less than 20nt in length discarded. Read alignments were performed with HISAT2 (Kim et al., 2019) using the HISAT hg38 index - an expansion of the hg38 reference genome to include information on Refseq transcripts and common SNPs. Alignment-free quantifications were performed for the Sam68 knockdown RNA-seq data, using Kallisto v0.42.4 (Bray et al., 2016) against GENCODE version 27 basic annotation transcripts. Kallisto sequence bias correction was turned on, and the default value of 100 bootstraps was used for estimates of quantification variance.

2.2 Statistical analysis and data visualisation

Unless otherwise stated, statistical analyses were performed using R v3.6.1 (R Core Team, 2019). Principal component analysis (PCA) was employed via the “prcomp” base R function, and linear modelling for PCA regression was performed with the “lm” base R function. Analysis and visualisation of intersections between gene or junction sets resulting from differential analyses was performed using the SuperExactTest R package (Wang et al., 2019). Heatmaps were created using the pheatmap R package v1.0.12 (Kolde, 2019). Visualisation of local splicing variations (LSVs) and junction PSIs was performed with the MAJIQ “voila” module (Vaquero-Garcia et al., 2016). The motifStack R package (Ou et al., 2018) was utilised for visualisations of PSSMs as sequence logos. Finally, analysis of receiver operating characteristic (ROC) area under the curve (AUC) was performed using the pROC R package (Robin et al., 2011). 95% confidence intervals of the AUC were calculated using a stratified bootstrapping approach according to default parameters of the pROC package ‘roc’ function. All other plots were generated using the ggplot2 R package (Wickham, 2012).

2.3 Differential splicing analysis

An initial comparison of several splicing analysis tools was performed: SplAdder (Kahles et al., 2016), SUPPA (Alamancos et al., 2015), VAST-TOOLS v1.1.0 (Tapial et al., 2017), and MAJIQ v1.1.7a (Vaquero-Garcia et al., 2016). SplAdder was run using the GENCODE v25 comprehensive reference annotation, with the highest confidence level of 3 used for splice event detection. SUPPA was used to define alternative splice events (specifically exon skipping, alternative 3'/5' splice site usage, mutually exclusive exon usage, or intron retention) from Kallisto derived transcript quantifications, prior to differential splicing analysis. For analysis with VAST-TOOLS, read data were aligned using the VAST-TOOLS aligner against the custom VAST database of human splicing variants; before quantification and differential splicing analysis was performed with VAST-TOOLS "count events" and "diff" functions. MAJIQ was selected for further use after initial comparisons with other splicing tools. The recommended reference genome - GRCh38.p8, was used in combination with RNA-seq sample data to define LSVs with the MAJIQ "build" module. MAJIQ differential splicing analysis was performed with the "dpsi" and "dvoila" modules using default parameters, such as for minimum numbers of quantifiable reads per sample for inclusion of each LSV.

2.4 Differential gene expression analysis

The Kallisto-Sleuth pipeline (Pimentel et al., 2017) was used for differential gene expression analysis, with the exception of the ENCODE data (see below). Kallisto (Bray et al., 2016) uses a pseudo-alignment approach allowing quantification directly from FASTQ files. Further, the speed of pseudo-alignment allows a bootstrapping procedure which facilitates estimation of quantification uncertainties. These estimates of uncertainty are incorporated as additional variance parameters in downstream differential gene expression analysis by the accompanying analysis tool Sleuth. The Kallisto-Sleuth approach was used for analysis of the Sam68 data (Chapter 5) and CpG HIV-1 data (Chapter 6), and no analysis of gene expression was performed in Chapter 4, where the focus was exclusively on splicing. For the analysis of ENCODE project knockdown RNA-seq (Chapter 3), pre-computed gene expression counts were obtained via the ENCODE portal (Davis et al., 2018) (ENCODE accessions in Appendix 8.1). The ENCODE analysis pipeline utilizes RSEM (Li and Dewey, 2011) for expression quantification, using STAR-aligned (Dobin et al., 2013) BAM files as input. For convenience, these pre-computed gene expression quantifications were used, rather than additionally running Kallisto on these ENCODE data.

Since Sleuth is designed for use with Kallisto-derived gene quantifications, and requires estimates of the uncertainty of quantification, an alternative differential gene expression analysis approach was employed for these data. To this end, limma (Ritchie et al., 2015) was applied for analysis of differential gene expression with these ENCODE samples, directly using the gene-level counts downloaded from the ENCODE portal. Benjamini-Hochberg correction was applied for calculation of all false discovery rates (FDR).

2.5 Gene ontology enrichment analysis

Gene Ontology (GO) enrichment analysis was performed against the "Biological Process" ontology (2019) with the gProfileR R package v0.7 (Reimand et al., 2018). Analysis was restricted to ontology categories with fewer than 100 gene members. The gene set counts and sizes (g:SCS) (Reimand et al., 2007) framework for multiple testing correction was used, and resulting significant terms were hierarchically filtered with the "moderate" stringency setting. Background gene sets were defined as genes expressed above a minimum threshold: for analysis of Sam68 knockdown data, a threshold of one transcripts per million (TPM) in at least one sample was used, for analysis of CD4+ T cell timecourse data, an alternative filtering strategy based upon quantifiable LSVs was used – see "Analysis of CD4+ T cell Activation and Polarisation Timecourse Data" section. To define related groups of enriched GO terms as depicted in the GO enrichment results figure (Figure 4-3), pairwise semantic similarity between significant terms was obtained via the GOSemSim R package v3.10 (Yu et al., 2010) using the "Wang" metric (Wang et al., 2007) based upon analysis of the GO graph structure. Hierarchical clustering was performed using these similarity scores, and terms were grouped with a tree cutting algorithm (Langfelder et al., 2008).

2.6 Splicing Motif Activity Response Analysis (S-MARA) workflow

2.6.1 Compilation of splicing factor motifs

In order to utilise a motif-based approach to study splicing, various splicing factor motifs in the form of PSSMs were compiled. RBPmap (Paz et al., 2014) is an RNA motif mapping tool which utilises a motif database consisting of 92 human motifs defined by Ray *et al.* through use of RNAcompete - a high-throughput *in vitro* assay (Ray et al., 2013), in addition to 26 motifs from other heterogeneous experimental sources. These PSSMs were obtained as part of the RBPmap v1.1 package (obtained from http://rbpmap.technion.ac.il/). These data were

combined with 131 motifs derived through a 2018 RNA-Bind-n-Seq (RBNS) study (Burge et al., 2018) (obtained through communication with author – Daniel Dominguez).

RBP motifs were filtered against a custom-defined list of 122 splicing factors. A 2014 census (Gerstberger et al., 2014) defined 1542 human RBPs, of which 692 were defined as primarily binding to mRNA targets. Of these, 247 genes were annotated with at least one of a number of defined GO "Biological Process" terms relating to roles in alternative splicing, as investigated through the biomaRt R package (Durinck et al., 2009). The selected splicing terms of interest were: "RNA splicing", "mRNA splicing, via spliceosome", "regulation of alternative mRNA splicing, via spliceosome", "regulation of RNA splicing", "mRNA splice site selection", "positive regulation of mRNA splicing, via spliceosome", "alternative mRNA splicing, via spliceosome", "regulation of mRNA splicing, via spliceosome", "positive regulation of RNA splicing", "mRNA 5'-splice site recognition", "mRNA 3'-splice site recognition", "pre-mRNA 3'-splice site binding", "negative regulation of RNA splicing", "pre-mRNA 5'-splice site binding", and "mRNA cis splicing". At this stage, 135 RBPs for which the major function is in regulating mRNA translation (i.e. translation initiation factors or poly(A) binding proteins) or cytoplasmic RNA stability were removed from the splicing factor list. Several additional splicing factors were then manually added to this list (ELAVL3, ELAVL4, HNRNPAB, HNRNPCL1, HNRNPDL, HNRNPUL2, MATR3, TIAL1, SRRM3, and EWSR1). The resulting set of 122 splicing factors is in Appendix 8.3.

Intersecting this splicing factor list with the set of RBP motif data resulted in 148 motifs representing binding preferences for 74 splicing factors (Appendix 8.3). This splicing factor-motif set contained a number of redundant motifs, including cases in which different experimental approaches produced highly similar but non-identical PSSMs for a given splicing factor. To address this redundancy, splicing factor-motifs were further processed with the MotIV v1.39.0 R package (Mercier and Gottardo, 2018). Initially, motif PSSM information content was calculated and trailing edges with < 0.3 bits of information were trimmed. Motifs were compared in a pairwise manner using the MotIV "motifDistances" function. In brief, PSSMs were aligned via the Smith-Waterman local alignment algorithm. Per nucleotide Pearson correlation coefficients were then computed from the aligned PSSMs, before taking the average correlation across the PSSM as a similarity measure for each pair of motifs. The resulting correlation coefficients were then used for hierarchical clustering with the "average" agglomeration method. Finally, the resulting distance tree was used to define clusters of

motifs as those with in-group distances (1 – average Pearson correlation coefficient) < 0.001. Motifs within each group were aligned and averaged per nucleotide position to generate consensus motifs via the MotifStack R package (Ou et al., 2018). This process resulted in a final set of 103 splicing factor motifs (see Results Figure 3-2).

2.6.2 Generation of motif count matrices

For the prediction of potential splicing factor binding sites, genomic regions of interest flanking splice junctions were first defined. MAJIQ quantifies relative splice junction usage as a percent selection index (PSI), which describes the relative usage of a single splice donor or acceptor with each of the possible pairing splice acceptor/donors in an LSV (Figure 2-1). Each PSI value therefore describes the usage of a fixed junction with one of several variable junctions (Figure 2-1). The RNA sequences flanking each variable junction potentially contain motifs that promote the binding of splicing factors and influence the relative usage of this splice junction, and thus its PSI. For each variable junction, these "regulatory regions" were defined as the area spanning 300 nt upstream-and-300 nt downstream of the junction (600 nt in total). This range corresponds with the region thought to contain the majority of SREs based upon previous investigations (Barash et al., 2010; Zhang et al., 2005). If a neighbouring exon boundary was less than 300 nt from a given variable junction, then it was used as the start/end point of that regulatory region instead (illustrated by the green splice junction in Figure 2-1). These regulatory regions were used to create FASTA sequence files of the corresponding hg38 genomic sequence via the "getfasta" function of BEDTools v2.25.0 (Quinlan, 2014). Finally, these FASTA files were scanned for the presence of potential splicing factor binding sites using the command line implementation of RBPmap v1.1.0 with our custom-defined set of 103 compiled splicing factor PSSMs. For each sequence, the cumulative number of motif matches with a probability < 0.001 relative to the default background model was calculated per motif using a sliding window approach. The resulting splice junction motif count matrix, in addition to the splice junction usage (PSI) matrix, was then provided as input to MARA.

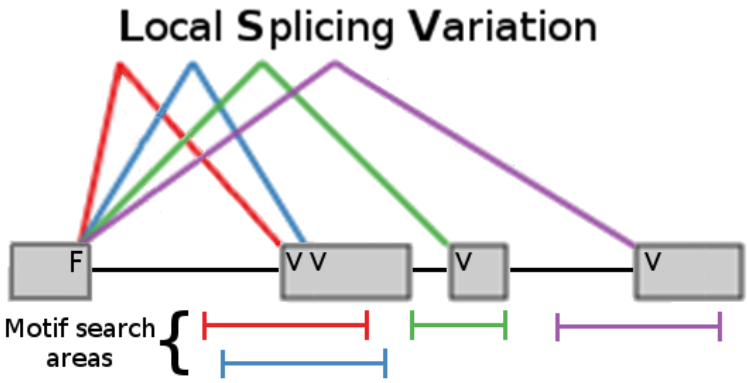

Figure 2-1. Schematic of an example local splicing variation and corresponding RNA regions scanned for the presence of splicing factor motifs. F = Fixed splice junction. V = variable splice junctions. Coloured motif search areas match the corresponding variable splice junctions. Red and blue variable junctions represent alternative 3' splice sites. Red, blue, and purple junction search areas are 600 nt long. The green junction has a smaller motif search area (regulatory region), representing a case where flanking exon boundaries are closer than 300 nt from the variable junction.

2.6.3 Motif Activity Response Analysis (MARA)

The Integrated Analysis of Motif Activity and Gene Expression Changes of Transcription Factors (IMAGE) (Madsen et al., 2018) implementation of MARA was employed (version 1.1 as obtained from https://github.com/JesperGrud/IMAGE). Specifically, R code for the calculation of motif activity and target prediction was extracted from the "Regression.R" script. The estimation of motif activity was performed as detailed by the IMAGE authors (Madsen et al., 2018). In brief, the motif matrix is first centred so that the mean count for each motif is zero. The PSI matrix is logit transformed (with the "logit" base R function) to create unbounded values from the set of 0-1 bounded PSI values, before per-sample centering and scaling is performed. Motif activities are calculated using ridge regression via the glmnet R package (Zou and Hastie, 2005). The use of motif count frequencies as predictors presents a problem of multi-collinearity, whereby similar motifs have similar count distributions. This in turn can present a problem of over-fitting when estimating motif activities. This problem motivates the use of ridge regression, in which an optimised regularization parameter (λ), favours smaller β coefficients and reduces variance in a variance-bias tradeoff. Optimal λ was determined through 10-fold cross-validation, whereby a 90%/10% split of junction data is iteratively used as a training/test set respectively.

The original MARA model centres on gene expression values. A representation of the model that is solved for PSI values is as follows:

Equation 2. Modelling splicing as a function of motif activities.

$$PSIj = \sum_{m} A_{s,m} \cdot N_{j,m}$$

PSIj is the PSI at a specified junction, *As,m* is the per-sample per-motif activity that is estimated via ridge regression, and *Nj,m* is the number of motif counts per-motif per-junction. Resulting motif activity values were zero-centred prior to downstream analysis.

IMAGE implements a "leave-one-out" analysis for inference of regulatory relationships between transcription factor-motifs and target genes. In this procedure, the above model is run in a reduced form in which the counts for the motif whose targets are to be identified are set to zero. Expressed as applied to splice junction data as herein, the difference in estimated *PSIj* accuracy (observed *PSIj* – estimated *PSIj*) between the full and reduced model is then calculated. Targets for each motif are then defined as those in which *PSIj* estimation accuracy is decreased in the full model, and for which motif counts are non-zero, with the decrease in estimation accuracy relative to the full model used as a score to represent the strength of association between each junction and motif. This procedure was applied for the prediction of splicing factor target splice junctions using the ENCODE knockdown data. Target splice junctions were predicted for each splicing factor motif using the above model applied to knockdown samples of the associated splicing factor along with all control samples, in a per-knockdown manner. This approached mirrored the analysis of splicing factor knockdown-induced differential splicing.

2.7 Motif enrichment analysis

Motif count enrichment analysis was performed against splice junction sets of interest defined through differential splicing analysis performed using MAJIQ. For each motif, count distributions were compared between junctions of interest and "background junction sets", which consisted of all other splice junctions with evidence for use in the RNA-seq data, but which were not in the "group of interest". One tailed Wilcoxon rank sum tests were used to test the probability of the null hypothesis that motif counts were not greater in the junction set of interest relative to background splice junctions. FDR correction was performed, and an FDR < 0.05 was taken as evidence that a given motif was over-represented ("enriched")

amongst splice junctions of interest, thus providing evidence of a potential regulatory role for the given splicing factor motif in the regulation of alternative splicing amongst those splice junctions.

2.8 Analysis of ENCODE project RNA-binding protein knockdown data

Table 2-1. ENCODE project shRNA-treated samples.

Cell Line	RBPs targeted via shRNA	Control shRNA-treated samples	Batches
K562	219	48	7
HepG2	225	50	7

Data describing shRNA knockdowns of 241 RBPs across two cell lines (HepG2 and K562) were obtained from the ENCODE data portal (Table 2-1). MAJIQ was applied to quantify per-sample LSVs genome-wide using pre-aligned BAM files (as available through the ENCODE portal), with default MAJIQ filters applied such as a minimum of three reads per-LSV per-sample. Additional filtering of LSVs was then performed, whereby only LSVs with sufficient read coverage to be quantifiable in at least 80% of samples per-cell-line were considered for downstream analysis with MARA. These ENCODE data were generated in 49 batches, and batch correction of PSI values was therefore performed via the "ComBat" function of the sva R package v3.32.1 (Leek et al., 2012). Analysis of significant changes in motif activity induced by RBP knockdowns was performed via Student's t-test followed by FDR estimation. Linear modelling for predictors of significant knockdown-induced changes in motif activity was performed using the base R "lm" function. Further, a metric for quantifying the knockdown-induced change in motif activity across all 103 splicing factor-motifs (global change in motif activity) was defined. To this end, the Euclidean distance between vectors of motif activity between all pairwise combinations of control samples with knockdown samples was calculated. The average of these distances was then used to represent the global change in splicing factor motif activities.

2.9 Analysis of CD4+ T cell activation and polarisation timecourse data

Table 2-2. CD4+ T cell timecourse samples

Condition	Biological donors	Time points (hrs)
None/naive	3	0
CD3 & CD28 stimulation plus Il-2 treatment (T_{h0})	3	0.5, 1, 2, 4, 6, 12, 24, 48, 72
CD3 & CD28 stimulation plus IL-2 & IL-4 treatment (T_{h2})	3	0.5, 1, 2, 4, 6, 12, 24, 48, 72

FASTQ files were obtained from the Sequence Read Archive (Table 2-1). After alignments, BAM files from technical replicates were merged with the Picard-tools v1.113 "MergeSamFiles" function (obtained from http://broadinstitute.github.io/picard/). For the per-sample application of MARA, the full set of quantifiable LSVs and corresponding splice junctions for each sample were used as input (i.e. numbers of input LSVs/junctions varied somewhat per sample). For other analyses which make simultaneous use of data from multiple samples, a consensus set of LSVs was defined as those with non-zero variance in PSI, and with sufficient read data to be quantifiable in at least 50% of samples using MAJIQ default parameters.

2.9.1 Definition of correlation modules from motif activities and junction PSIs

In order to identify distinct profiles of temporal regulation across the timecourse of CD4+ T cell activation, a module analysis in the form of Weighted Gene Co-expression Analysis (WGCNA) (Langfelder and Horvath, 2008) was applied. WGCNA uses a correlation-based approach, and takes as input a correlation matrix which describes, for instance, the correlations of gene activities across a timecourse. These correlations are scaled by an optimized parameter in order to transform the topology of the resulting network to be 'scale-free', an observable property of many biological networks (Zhang and Horvath, 2005). Subsequently, modules of highly correlated genes can be identified through a hierarchical clustering-based approach, with the resolution of these modules adjustable according to several user-determined parameters. WGCNA is generally applied to gene expression estimates, but can equally be applied to other measures of gene activity such as splicing (Iancu et al., 2015).

Motif activity was calculated through MARA as described. Motif activity values were then hierarchically clustered via average linkage of Euclidean distances, and the resulting distance tree was used to define modules with the "cutreeDynamic" algorithm (Langfelder et al., 2008) of the WGCNA R package. Minimum module size was set to three motifs, and the "deepSplit" parameter which determined sensitivity to cluster splitting, was also set to three. As the dimensionality of the splice junction PSI data was much greater, the full WGCNA workflow was employed via the "blockWiseModules" function to define co-splicing modules of splice junctions. The "signed hybrid" network type was used so that negatively and positively correlated junctions were grouped into separate modules. Initial module definition was performed with the "deepSplit" parameter set to 3, and a minimum module size of 30 junctions. Modules were summarised using the first principal component and, in keeping with WGCNA nomenclature, these values are referred to as eigenJunctions or eigenMotifs, for modules of junctions or motifs respectively. Pairs of modules with an eigenvalue Pearson correlation of >= 0.75 were merged prior to further analysis.

2.9.2 Statistical analysis of motif activity and junction PSI

To model relationships between time after activation in CD4+ T cells and variables of interest (junction PSI, motif activity, eigenJunctions, and eigenMotifs), linear mixed effect spline modelling was performed via the lmms R package v1.3.3 (Straube et al., 2015). Null intercept-only models were contrasted with full models in which coefficients for time-after-activation and cell type (T_{h0}/T_{h2}), or cell type-time interactions terms were fitted, allowing the significance of these experimental parameters in explaining the variables of interest to be assessed. Naïve CD4+ T cells were excluded from these statistical analyses since only a single replicate per donor (time point zero) was available for this condition.

2.10 Sam68 knockdown experimental procedures

Table 2-3. Samples used in Sam68 knockdown experiment.

shRNA target	Activation status	Biological donors
Sam68 (shRNA 1)	Both resting & activated	3
Sam68 (shRNA 2)	Both resting & activated	3
Both resting & activated	Both resting & activated	3
None/untransduced	Both resting & activated	3

N.B. Experimental work detailed below was performed by Laura Hidalgo.

CD4+ T cells were isolated from peripheral blood mononuclear cells from three donors before *in vitro* culture and activation via CD3/CD28 stimulation to facilitate subsequent transductions (day 0). At day 1, cells were either 1) left untransduced as a control, 2) transduced with a control scrambled shRNA, or 3) transduced with one of two shRNAs targeted to the Sam68 mRNA. The ΔLNGFR transduction tag system (Lauer et al., 2000) was used, which facilitated isolation of successfully transduced cells at day 5. After a further three days (at day 8), RNA was extracted from cells from each of the three conditions, which are now in a resting state having not been exposed to the activation stimulus for a number of days. On the same day, cells were re-activated via further CD3/CD28 stimulation, before final isolation of RNA from these re-activated cells a day later (day 9). RNA samples were then used for poly(A) selected RNA-seq using an Illumina TruSeq stranded mRNA library preparation kit following the manufacturer's protocol. This resulted in a total of 24 RNA-seq libraries – 4 treatment conditions (untransduced, control shRNA, Sam68 shRNA 1, Sam68 shRNA 2), 2 activation states (resting, re-activated), and 3 donors (biological replicates).

Analysis of resulting RNA-seq libraries was performed as described in specific Methods subsections – "RNA-seq pre-processing", "Statistical analysis and data visualization", "Differential splicing analysis", "Differential gene expression analysis", "Gene ontology enrichment analysis", and "Motif enrichment analysis".

2.11 Analysis of the effects of introducing CpG dinucleotides to the HIV-1 genome

Table 2-4. Samples used for analysis of HIV-1 CpG content.

Replicates	HIV construct used to transfect HeLa Cells
2	pHIV-1_{NL4-3}/wildtype
2	HIV-$1_{gag22-165}$CM
2	HIV-$1_{gag22-261}$CM
2	HIV-$1_{gag22-378}$CM

2.11.1 Introduction of CpG dinucleotides to the HIV-1 genome – experimental procedures

N.B experimental work detailed below was performed by Irati Antzin-Andeutza and Mattia Ficarelli. Detailed further in (Ficarelli et al., 2020).

The pHIV-1_{NL4-3} plasmid, which contains the HIV-1 proviral sequence from the pHIV-1_{NL4-3} isolate (Adachi et al., 1986) was used for production of infectious HIV-1. Modified viral constructs HIV-$1_{gag22-165}$CM, HIV-$1_{gag22-261}$CM and HIV-$1_{gag22-378}$CM were used, which have the designated sequences from the patented pHDMHgpm2 vector (Gray et al., 2005). These constructs contain modifications to the *gag* region which have increased the number of CpG dinucleotides whilst preserving coding sequence (codon modification). Additionally, a construct with CpGs introduced into *env* was used, HIV-$1_{env88-561}$CM, which was produced as detailed in (Ficarelli et al., 2019)**.** These viral constructs were used to transfect HeLa cells in order to generate infectious viral particles.

For the production of ZAP knockout cell lines by CRISPR-Cas9, ZAP-targeting guide sequences were inserted into a lentivirus-based CRISPR plasmid. These plasmids were used to transfect HEK293T cells, and virus-containing supernatant was then harvested 48 h after transfection, and used to transduce HeLa cells in order to deplete ZAP.

For analysis of the relative production of viral proteins and RNA, HeLa cells were transfected with either wild-type or CpG codon modified plasmids. The cells were lysed 48 h post-transfection, and the medium was recovered for analysis of either proteins or RNA. To analyse HIV-1 protein abundance, virions were pelleted via centrifugation. The resulting pellets were then used as substrate for immunoblotting against HIV-1 $p24^{Gag}$, gp160/120, or Hsp90 as a control. To analyse RNA abundance, RNA was extracted from media after cell lysis. Quantitative PCRs (qPCRs) were performed in triplicate. Genomic RNA was quantified using primers which specifically amplify the full-length unspliced RNA, whilst total RNA was quantified using primers which amplify all HIV-1 transcripts whether spliced or unspliced. To generate RNA-seq libraries, the Illumina TruSeq stranded mRNA library preparation kit was used, before sequencing was performed with the Illumina HiSeq instrument.

The TZM-bl cell infectivity assay (Sarzotti-Kelsoe et al., 2014) was used to measure relative infectious viral particle production. To this end, supernatant was recovered from HeLa cells 48 h post-transfection and used to infect TZM-bl cells overnight. Forty-eight hours post-infection, the cells were lysed, and the amount of infectious-virus production was measured as relative light units per second after induction of β-galactosidase using the Galacto-Star system.

2.11.2 Analysis of RNA-seq libraries from HeLa cells transfected with CpG modified HIV-1 viruses

RNA-seq reads were aligned to the human genome (hg38) and HIV-1 NL4-3 genomic RNA simultaneously using Hisat2. HIV-mapping junction spanning reads were isolated using regtools (Feng et al., 2018) to allow per-junction read counting. To visualise read data spanning specific splice junctions of interest, data from replicates were first merged using the Picard (http://broadinstitute.github.io/picard) "MergeSamFiles" function, before sashimi plots were generated using the Gviz R package (Hahne and Ivanek, 2016). Other aspects of the analysis were performed as described in specific Methods subsections: "Statistical analysis and data visualization", and "Differential gene expression analysis".

Table 2-5. Data sources in this study.

Description	Associated paper/s	Source	Study accession	Sample accessions
PSSMs from RBPmap database	(Paz et al., 2014)	http://rbpmap.technion.ac.il/	NA	NA
RBNS-derived PSSMs	(Burge et al., 2018)	Personal communication with Daniel Dominguez	NA	NA
ENCODE project RBP-knockdown, RNA-seq, BAM files, and gene quantifications	(Burge et al., 2018), (Nostrand et al., 2018)	ENCODE portal (Davis et al., 2018)	-	See appendix 8.1

CD4+ T cell activation and polarisation timecourse RNA-seq	(Henriksson et al., 2019)	Sequence Read Archive (Leinonen et al., 2011)	ERP105662	See appendix 8.2
CD4+ T cell activation, single time point, RNA-seq	(Ni et al., 2016)	Sequence Read Archive (Leinonen et al., 2011)	SRP058500	SRX1033297, SRX1033298, SRX10834928, SRX1835120
Timecourse of CD4+ T cell activation in both memory and naïve cells	(LaMere et al., 2016)	Sequence Read Archive (Leinonen et al., 2011)	PRJNA296380	SRS1074588, SRS1074587, SRS1074585, SRS1074584, SRS1074579, SRS1074578, SRS1074586, SRS1074577, SRS1074613, SRS1074612, SRS1074611, SRS1074610, SRS1074605, SRS1074604, SRS1074603, SRS1074602

Chapter 3. Assessing the Performance of Motif Activity Response Analysis (MARA) Applied to Splicing Factor Biology

3.1 Introduction

MARA is a promising approach for inferring drivers of pre-mRNA splicing regulation within a given biological system. In order to apply MARA to the inference of splicing factor behaviour (S-MARA), a workflow must be implemented to generate the appropriate splicing-based input. When MARA is utilised to identify putative regulatory transcription factors, two sources of data are used: 1) genome-wide quantitative measurements (e.g. of gene expression), and 2) a matrix of matched motif counts representing potential regulatory factor binding sites (i.e. counts of transcription factor motif occurrences). Thus, to adapt MARA for analysis of alternative splicing, we have implemented a workflow to quantify genome-wide splicing variation from RNA-seq data, and to quantify splicing factor motif occurrences in RNA sequences flanking splice junctions.

Additionally, MARA has been optimised and validated for the purposes of inferring transcription factor motif activity (Balwierz et al., 2014; Madsen et al., 2018; The FANTOM Consortium et al., 2009), but not splicing factor motif activity. Whilst there is an analogy in how both splicing and gene expression are regulated through the actions of proteins towards *in cis* motifs, there are also several differences between these processes which may be relevant to MARA. A pertinent example is that splicing factor motifs are often shorter; with lower information content than transcription factor motifs (Burge et al., 2018; Madsen et al., 2018). In light of this, we assessed the ability of MARA to infer changes in splicing factor motif activities which reflect differential splicing across biological conditions.

Recent work conducted through the ENCODE project has generated several hundred RBP-knockdowns across HepG2 and K562 cell lines (Burge et al., 2018; Nostrand et al., 2018). This work provides an excellent resource for investigating the functional contribution of RBPs to gene expression and RNA regulation, and a number of the shRNA-induced knockdown targets were splicing factors. As such, these data present an opportunity to apply MARA to samples in which splicing factor activity has been experimentally altered through a reduction in gene expression. Further, these data represent splicing factors with a range of RNA binding

preferences and motifs. The ability of S-MARA to link differential splicing with specific pre-defined regulatory motifs was assessed. S-MARA was compared to a motif enrichment analysis approach, which is a commonly applied method for inference of regulatory splicing factors (Carazo et al., 2018; Chen et al., 2014; Sebestyén et al., 2016).

The motif enrichment strategy is primarily chosen here to provide a baseline comparison for S-MARA. However, assessing motif enrichment as a procedure for inference of regulatory splicing factors is also of interest in itself. To our knowledge, this is the first formal assessment of the performance of a splicing-based motif enrichment analysis in identifying regulatory splicing factors.

3.2 Aims

To understand the full gene expression programme underlying a given biological process requires knowledge of context-specific regulatory splicing factors. I propose that application of MARA to the analysis of RNA-seq data will allow inference of such regulatory splicing factors. Splicing MARA could be applied to any biological system in which alternative splicing is of interest. I therefore aim to:

1. Implement a data processing workflow to generate splicing-based input matrices for MARA.
2. Assess the ability of MARA to infer changes in splicing factor motif activity induced through knockdown of splicing factors.
3. Contrast results from S-MARA with a motif enrichment approach based on differential splicing analysis followed by motif enrichment testing.

3.3 Results

3.3.1 Preliminary investigation of splicing analysis tools and compilation of splicing factor motifs

3.3.1.1 Comparison of RNA-seq differential splicing analysis tools

An ever-increasing number of approaches for RNA-seq based differential splicing analysis exist. Since there is not an established gold standard *per se* (Carazo et al., 2018; Ding et al., 2017; Hooper, 2014), several tools which employ different methodologies were selected for

comparison in a pilot study (Table 3-1). Several published RNA-seq data sets from studies of primary CD4+ T cell activation were utilised for this purpose. Ni *et al*. stimulated human CD4+ T cells via CD3/CD28 and isolated RNA for sequencing both prior to activation and at 18 hours post-activation (Ni et al., 2016). LaMere *et al*. isolated naïve and memory CD4+ T cells from PBMCs and activated them via CD3/CD28 stimulation in combination with IL-2, before performing RNA-seq prior and at 1, 5, and 14 days post-activation (LaMere et al., 2017). Differential splicing of CD45 upon CD4+ T cell activation, or in memory relative to naïve cells, was utilised as a positive control splicing event. Specifically, the ability of the selected analysis tools to detect exclusion of exon 4 in the activated/memory state was assessed. Expression of the isoform containing exon 4 (CD45RA), is used as a marker of the naïve state, such as by LaMere *et al*. in their study (LaMere et al., 2017).

Table 3-1. Features of differential splicing analysis tools selected for comparison.

Tool	Approach to defining splicing events	Splicing Model
MAJIQ (Vaquero-Garcia et al., 2016)	Reference genome plus RNA-seq evidence	Local splicing variants (LSVs) of arbitrary complexity
SplAdder (Kahles et al., 2016)	Splicing graph constructed using reference genome and augmented with RNA-seq data	Exon-based
SUPPA (Alamancos et al., 2015)	Splicing events defined via comparison of reference transcriptome isoforms	Isoform/Exon-based
VAST-TOOLS (Tapial et al., 2017)	Uses custom splice junction database derived through integrating reference genome with diverse expression data	Exon-based

CD45 splicing events were not reported by SplAdder or SUPPA. The version of SplAdder at the time of analysis (version 1) pre-processes a reference transcriptome and removes any genes which partially share genomic co-ordinates (e.g. overlapping sense-anti-sense transcripts) to prevent issues associated with assigning aligned RNA-seq reads to specific splice junctions in such instances. This approach is undesirable as it seems to be overly conservative, excluding many genes of interest. SUPPA defines exon skipping events as cases in which a single cassette

exon differentiates transcript isoforms. Splicing at the CD45 locus exceeds the simplicity of this model and is thus not reported by SUPPA. Note that newer versions of both SplAdder and SUPPA have been released since this analysis was performed. MAJIQ and VAST-TOOLS successfully detected preferential exclusion of CD45 exon 4 in memory CD4+ T cells relative to naïve cells in the LaMere *et al.* dataset (Figure 3-1). However, neither tool reported a change to exon 4 splicing at 18h post-activation in the Ni *et al.* dataset (Figure 3-1), which may be due to the splicing kinetics of this event. Both VAST-TOOLS and MAJIQ coincidentally identified 1998 genes as differentially spliced at 18 hrs post-activation of CD4+ T cells, with 634 of these genes being identified by both tools. This is a significant overlap (p = < 0.001) as assessed via hypergeometric test using a background of all genes having sufficient read coverage to be analysed by either tool [9076]). Therefore, both VAST-TOOLS and MAJIQ were deemed appropriate analysis tools producing partially overlapping results. MAJIQ defines and quantifies splicing events of arbitrary complexity, whereas VAST-TOOLS focuses on analysis of classically defined splicing events and could thus be considered more limited in this capacity. On this basis, MAJIQ was selected for all further analyses.

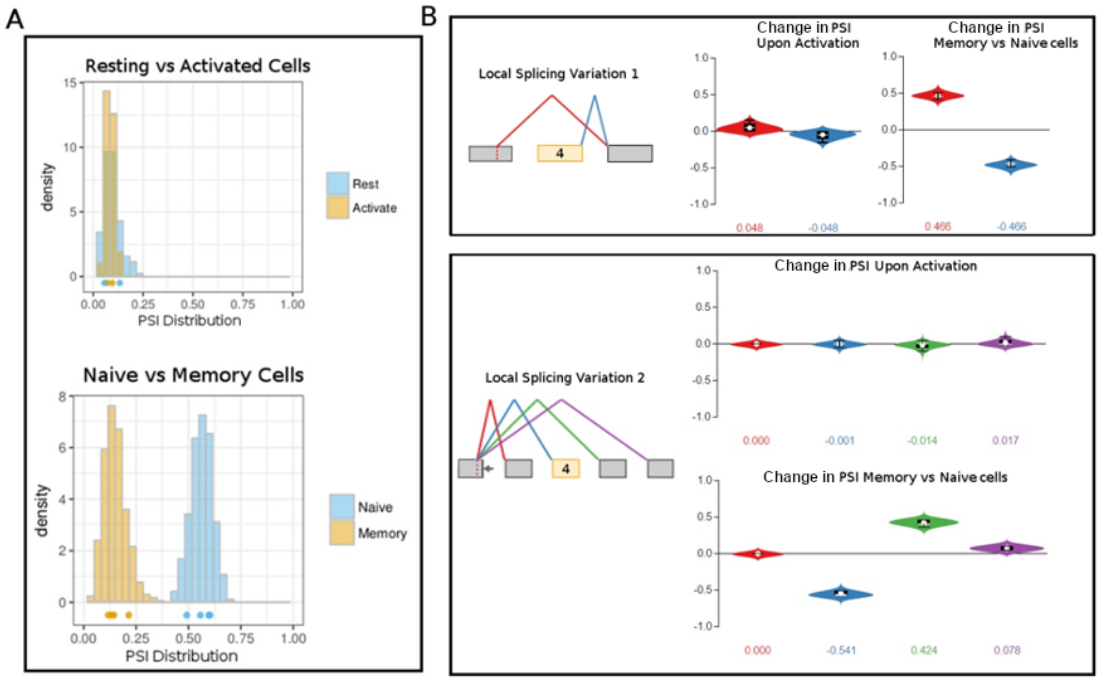

Figure 3-1. Splicing of CD45 exon 4 in different CD4+ T cell states. (A) Results reported by VAST-TOOLS. **(B)** Results reported by MAJIQ. Exon 4 was represented by MAJIQ as two LSVs. Violin plots show the distribution of the estimated PSI for the colour-matching junction in each

LSV schematic. Data comparing resting with activated CD4 T cells from (Ni et al., 2016). Data comparing naïve with memory CD4+ T cells from (LaMere et al., 2016).

3.3.1.2 Splicing factor motifs

A set of 103 splicing factor-motifs, which collectively capture the binding preferences of 74 splicing factors, was compiled for downstream application of MARA (Figure 3-2; see Chapter 2 for details of motif sources). Many splicing factors were represented by multiple motifs, and the median number of motifs for a given splicing factor was two. In cases where several motifs from different splicing factors were highly similar, these motifs were represented by a single consensus position specific scoring matrix (PSSM) (see Chapter 2 for details on consensus motif generation).

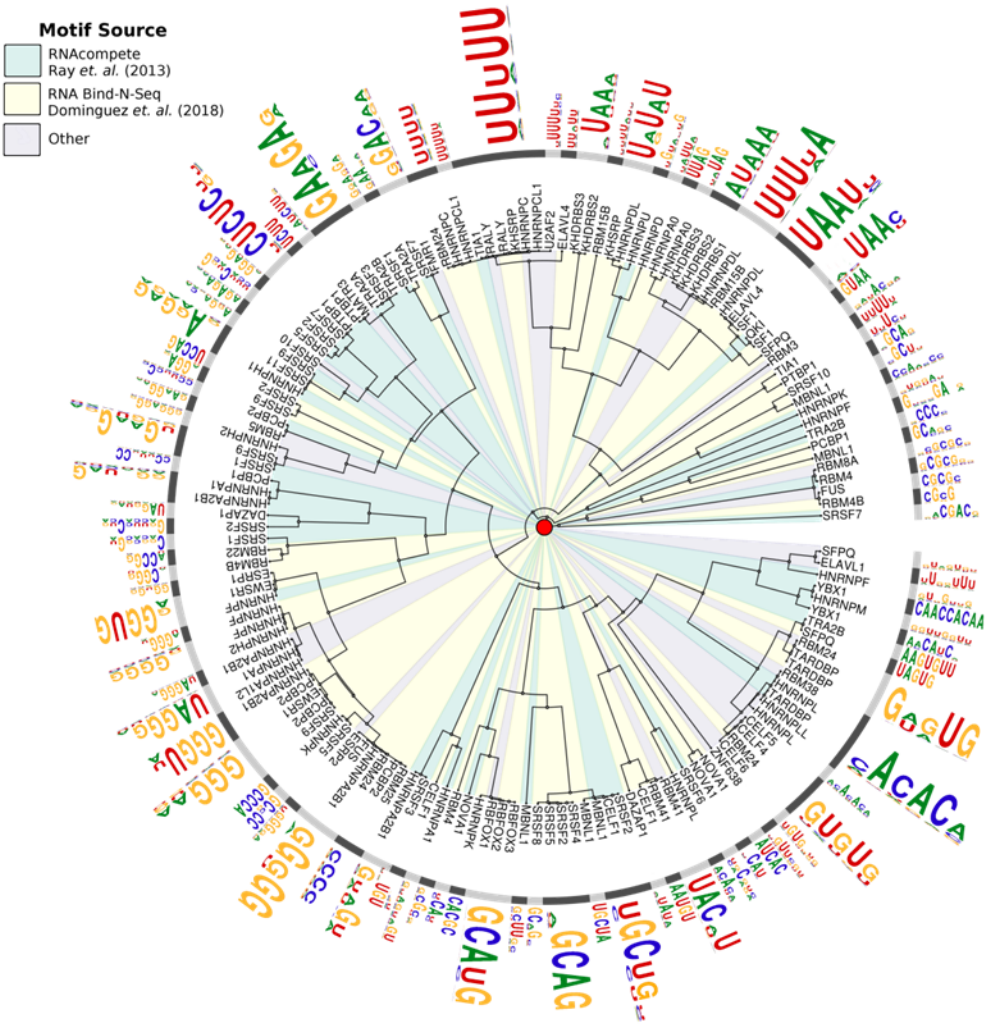

Figure 3-2. Splicing factor motifs utilised for S-MARA. Motifs are clustered according to Pearson correlation of the PSSMs after Smith-Waterman local alignments, and similar motifs have been merged to generate a consensus.

These motifs ranged in length from 4 to 9 nt (median = 5 nt), with an information content ranging from 3.77 to 18 bits (median = 7.52 bits). To allow comparison with the characteristics of transcription factor DNA motifs, the database of the IMAGE motif activity analysis tool (Madsen et al., 2018) was used, which contains a total of 1579 transcription factor motifs. The median transcription factor motif length was 12 nt, and the median information content was 12.104 bits, significantly greater than for the splicing factor motifs ($p < 2.2 \times 10^{-16}$ Wilcoxon rank sum test) (Figure 3-3). This highlights an important difference between DNA and RNA binding motifs which could potentially influence the application of MARA to splicing factor biology.

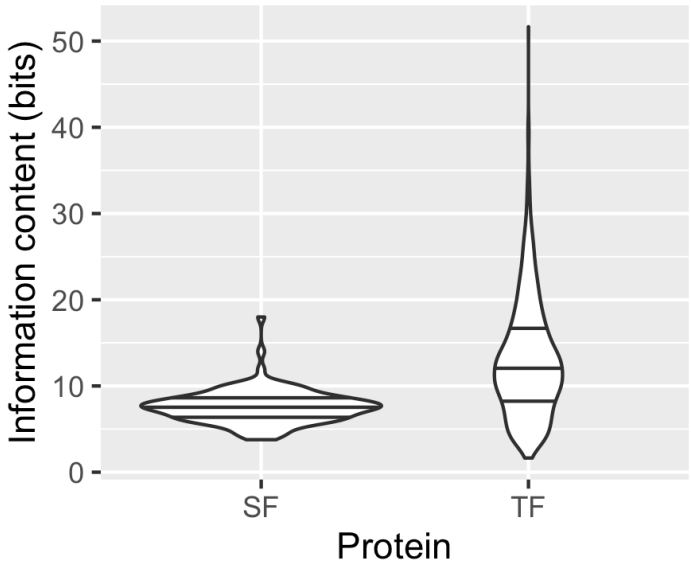

Figure 3-3. Information content of motifs associated with splicing factors or transcription factors. SF = splicing factor, TF = transcription factor. Data for 1579 TF motifs from the IMAGE database (Madsen et al., 2018) and the 103 SF motifs compiled for this study.

3.3.2 Pre-processing and quality control of ENCODE RNA-seq data

3.3.2.1 Effects of RBP knockdown on splicing in HepG2 and K562 cell lines

The ENCODE project shRNA knockdowns were conducted across 49 batches, with each batch containing two non-specific shRNA treated controls and a variable number of duplicate

knockdown samples in which different RBPs were targeted. A total of 48 control and 438 knockdown samples in K562 cells, and 50 control and 450 knockdown samples in HepG2 cells were used, totalling knockdowns for 241 RBPs. To determine the magnitude of depletion in each experiment, differential gene expression analysis was used to compare knockdown and control samples. This revealed that the RNA abundance of the RBP of interest was significantly reduced (FDR < 0.05) in 210/219 knockdowns in K562 cells, and 222/225 in HepG2, with a mean reduction in expression of 47% (Figure 3-4). Of the 241 RBPs analysed in the ENCODE dataset, 38 of these were present in our compiled splicing factor motif set (Figure 3-2), represented by a total of 67 different motifs. The majority of these splicing factors had a significant reduction in gene expression, with the exception of TRA2A in K562 cells and SRSF3 in HepG2 cells. The mean reduction in gene expression of splicing factors specifically was 45.6%. The knockdown data for these 38 RBPs (33 depleted in both cell lines, five in just one cell line) were selected for motif activity analysis.

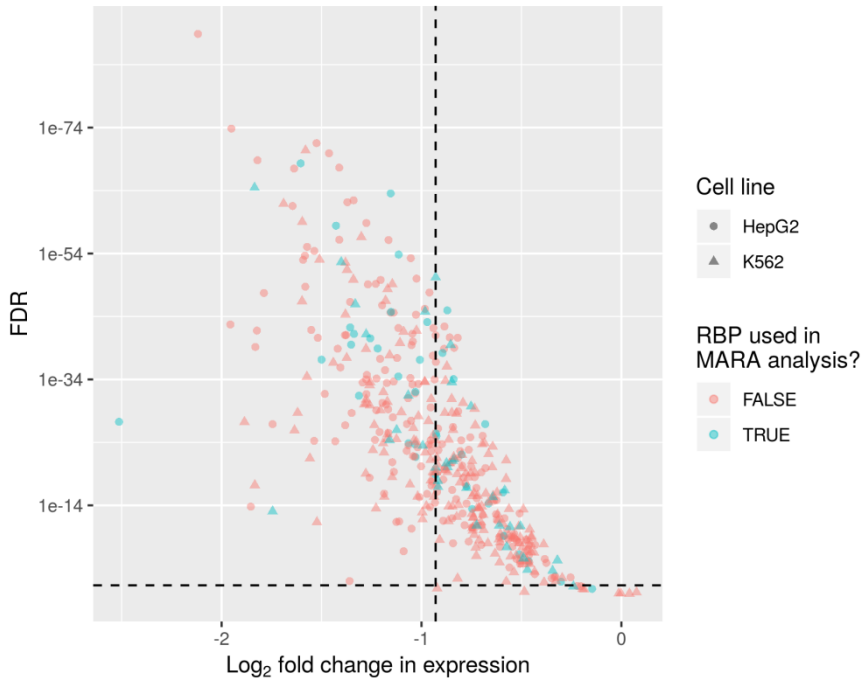

Figure 3-4. Volcano plot of RNA binding protein knockdowns in HepG2 and K562 cells. Horizontal and vertical lines mark FDR of 0.05 and the mean $\log_2$ fold change respectively. RNA-binding proteins annotated as splicing factors and having associated motif data are indicated as those being used for the MARA-based analysis. FDR is plotted on a $-\log_{10}$ scale.

The pool of splice junctions considered for analysis was defined as those with sufficient read coverage (see Chapter 2) in at least 80% of samples for each cell line independently, resulting

in ~116,000 junctions in HepG2 cells and ~113,000 junctions in K562 cells. The number of splice junctions with a significantly altered PSI after RBP-knockdown varied widely, with splicing factors affecting more junctions on average than non-splicing factor RBPs, as expected ($p = 1.6 \times 10^{-13}$ - Wilcoxon rank sum test) (Figure 3-5). Knockdown of *HNRNPC* in HepG2 cells had the largest effect on splicing, whilst *AARS* knockdown in K562 cells had the smallest overall effect (Figure 3-6). The magnitude of the knockdown-induced splicing effect was overall similar between HepG2 and K562 cells, with a Pearson correlation of 0.61 between numbers of differentially spliced junctions per knockdown. However, several cell-type specific effects were present. For instance, focusing specifically on splicing factors (Figure 3-7), *HNRNPC* and *HNRNPA2B1* knockdown had more pronounced effects in K562 cells, whilst *U2AF2* and *HNRNPK* knockdown had larger effects in HepG2 cells.

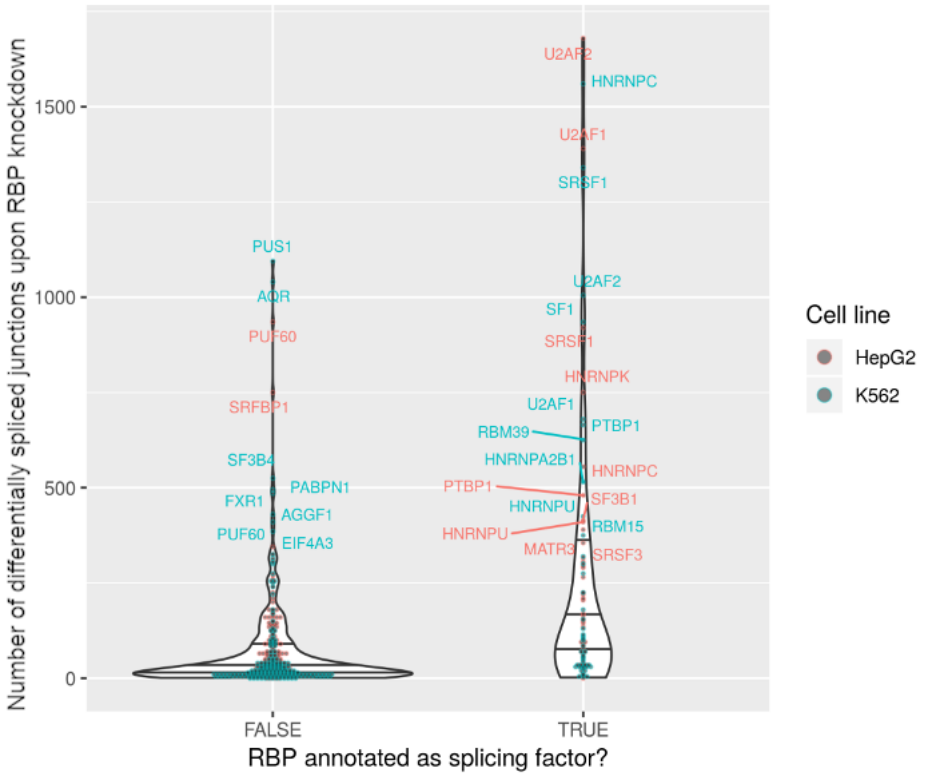

Figure 3-5. Effects of RNA-binding protein knockdowns on splicing in HepG2 and K562 cells. Knockdowns producing altered splicing of more than 300 splice junctions are labelled with the target RNA-binding protein.

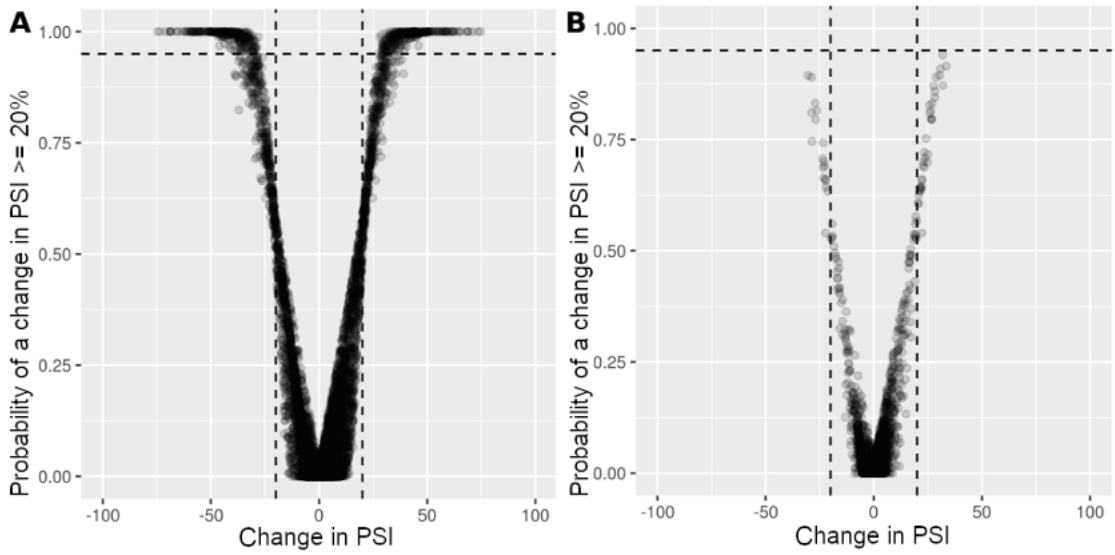

Figure 3-6. Volcano plots of RNA-binding protein knockdowns. (A) *HNRNPC* knockdown in HepG2 cells. **(B)** *AARS* knockdown in K562 cells. The horizontal lines mark 95% probability of a change in PSI >= 20%. Vertical lines mark -20%, and 20% change in PSI.

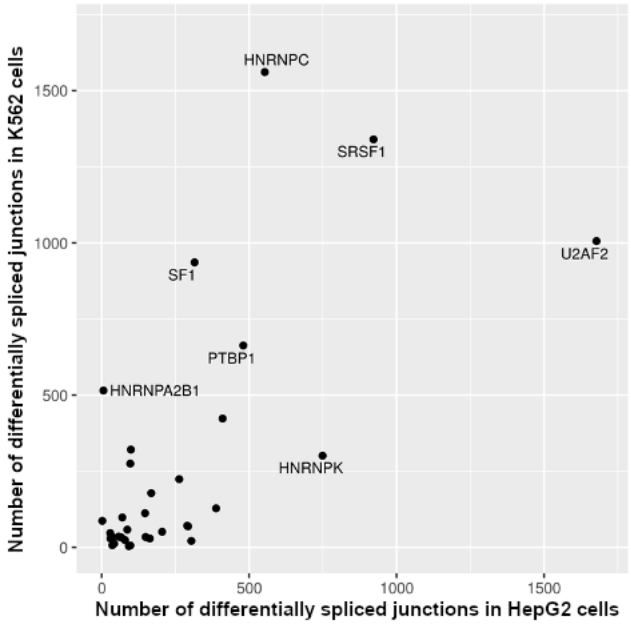

Figure 3-7. Effect of splicing factor knockdown in HepG2 or K562 cells.

The numbers of differentially spliced junctions upon knockdown of splicing factors are shown. Data relating to select splicing factors are labelled.

3.3.2.2 ENCODE data batch adjustment

Since ENCODE RBP-knockdown samples were processed in batches, potential batch effects were investigated. PCA analysis of genome-wide splice junction PSI values separated samples by batch along the first two PCs (Figure 3-8A). Batch effect adjustment of the PSI matrix via ComBat (Leek et al., 2012) removed the batch effect visually as assessed via PCA (Figure 3-8B). Comparing samples across batches is desirable since this allows greater numbers of control samples to be included in analysis of each knockdown, increasing statistical power. However, in the ENCODE study, RBP-dependent splicing effects were confounded with batch. Although each batch contained a common experimental condition in the form of control shRNA-treated samples, RBP-knockdown samples for a given RBP were always confined to a single batch. This characteristic of the study design may limit the ability to estimate and remove batch effects whilst adequately preserving the signal associated with the condition of interest (RBP-knockdown). To address this issue, both batch-adjusted and un-adjusted data were utilised for initial stages in the analysis, to assess potential downstream influence of the batch effect.

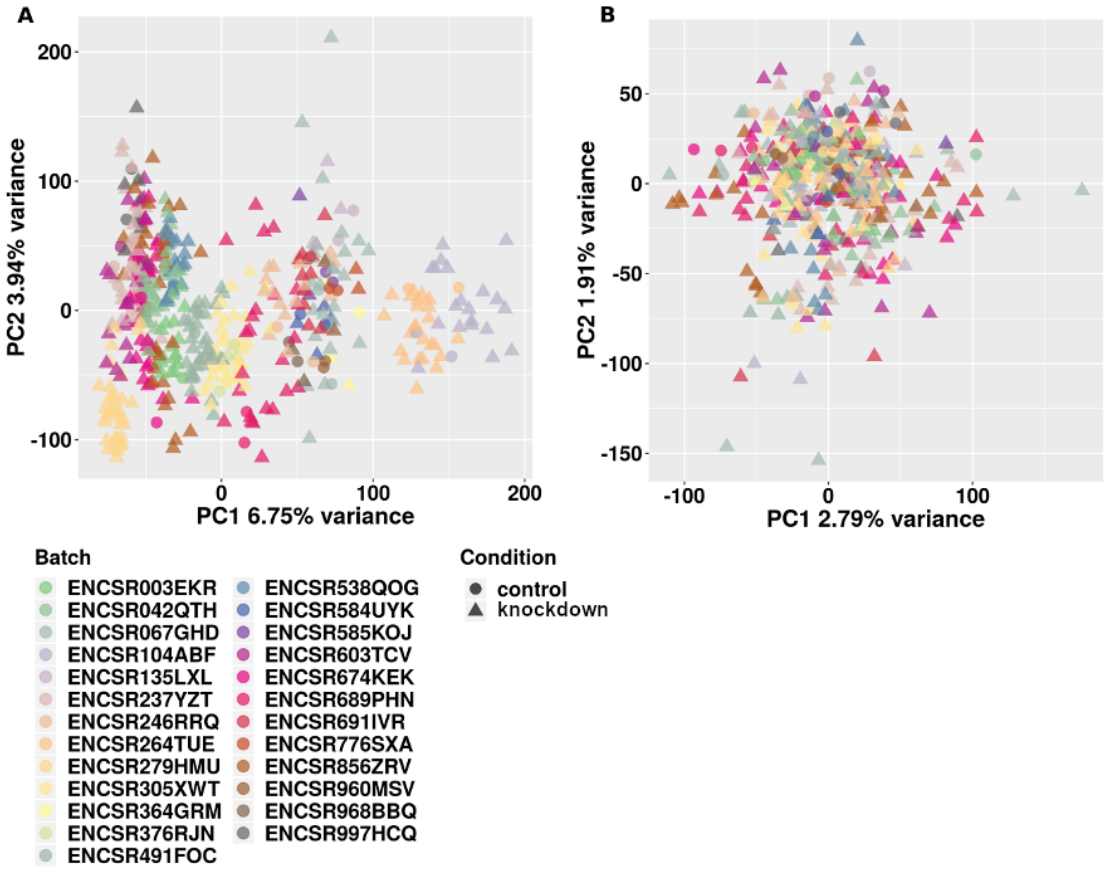

Figure 3-8. PCA analysis of genome-wide splice junction PSI values before and after batch correction. Data shown for HepG2 cells only, similar patterns observed for K562 cells. Batch

refers to the ENCODE experiment accession identifier. **(A)** Unadjusted data. **(B)** Data after batch correction via ComBat.

To assess the effect of RBP-knockdown on splicing factor motif activity, the change in motif activity between control and knockdown samples was defined as the difference between the mean motif activity in all controls and that in RBP-knockdown sample duplicates. As some RBPs have multiple associated motifs, this was performed for each motif associated with such RBPs. The change in motif activity values were highly correlated whether using batch-adjusted or unadjusted PSI values (r = 0.766) (Figure 3-9). Since changes in motif activity are the primary outcome of interest, and to avoid potential issues of batch-adjustment with an unbalanced experimental design, the unadjusted data were selected for further use.

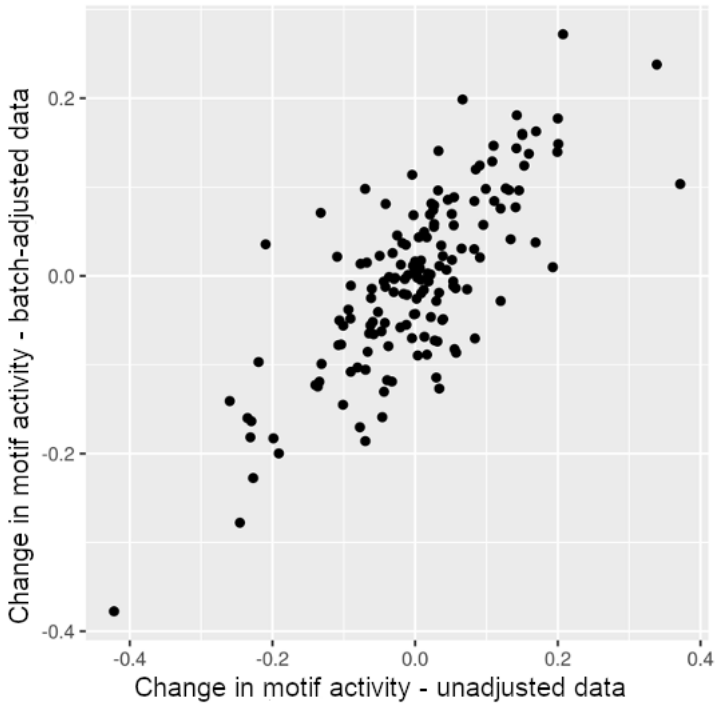

Figure 3-9. Effect of batch correction on motif activity changes upon splicing factor knockdown. Changes in splicing factor motif activities after splicing factor knockdown are shown.

3.3.3 Assessing the performance of MARA in identifying changes in splicing factor motif activity

3.3.3.1 Splicing factor knockdown induced motif activity changes

Of the 38 analysed splicing factor-knockdowns, 22 resulted in a change in activity for at least one of the splicing factor-associated motifs (FDR < 0.1). This corresponded to 17 splicing factors in HepG2 cells and 11 in K562 cells (Figure 3-10). Six splicing factors had a motif with significantly altered activity in both HepG2 and K562 cells, which is a non-significant overlap as assessed via hypergeometric test (probability of an intersection of this size or greater = 0.34) (Figure 3-11). In total, this equated to 25 motifs with a significant change in motif activity - 16 in HepG2 cells and 13 in K562 cells, with 4 common to both cells. The two splicing factors for which shRNA treatment did not significantly reduce gene expression (*SRSF3* in HepG2 cells and *TRA2A* in K562 cells) did not have a significant change in motif activity. However, overall the knockdown efficiency, as assessed via fold change reduction in expression, was not associated with significance of the change in motif activity (Figure 3-12A). Similarly, the number of splice junctions with differential usage upon splicing factor-knockdown was not related to significance of the motif activity change (Figure 3-12B). However, when considering the global change in activity across all 103 splicing factor motifs in controls relative to knockdown samples, an association with the number of differentially spliced junctions was present (Figure 3-13) (Pearson correlation, $r = 0.69$, $p = < 2.2 \times 10^{-16}$). In particular, some splicing factor-knockdowns with large effects on alternative splicing, but without significant changes in motif activity for their associated motif, such as *U2AF2*, did result in large changes overall across the 103 tested motifs (Figure 3-13). Thus, whilst not all splicing factor knockdowns resulted in detected changes in associated motif activity, MARA is able to report motif activity changes in proportion to the magnitude of splicing changes between samples.

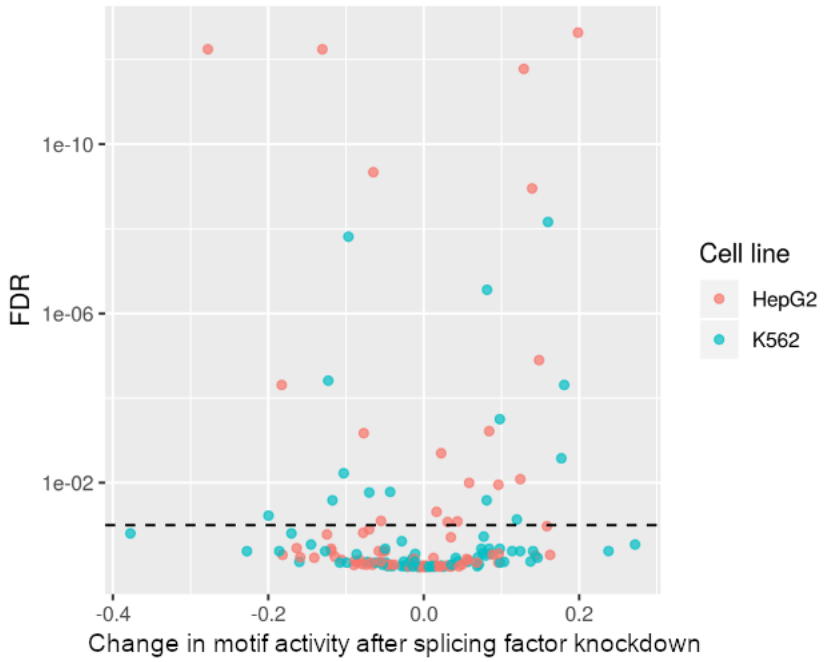

Figure 3-10. Volcano plot of motif activity changes upon splicing factor knockdown.

Horizontal line highlights FDR of 0.1. FDR is plotted on a -$\log_{10}$ scale.

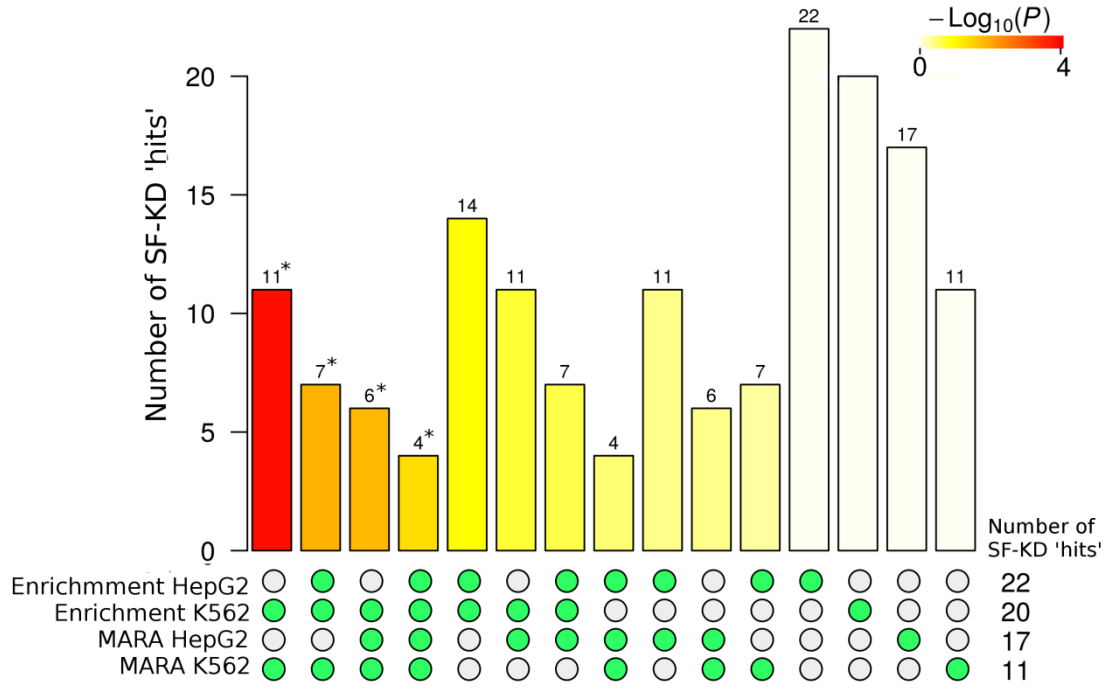

Figure 3-11. Intersections between splicing factor knockdowns with associated motifs identified through either a MARA-based or a motif enrichment-based approach.

"Enrichment" refers to motif enrichment analysis. Green circles indicate the analysis types for which the intersection number in the above bar chart refers to. p values are derived from

Fishers' exact test. * indicates a significant intersection ($p < 0.05$). SF = splicing factor, KD = knockdown.

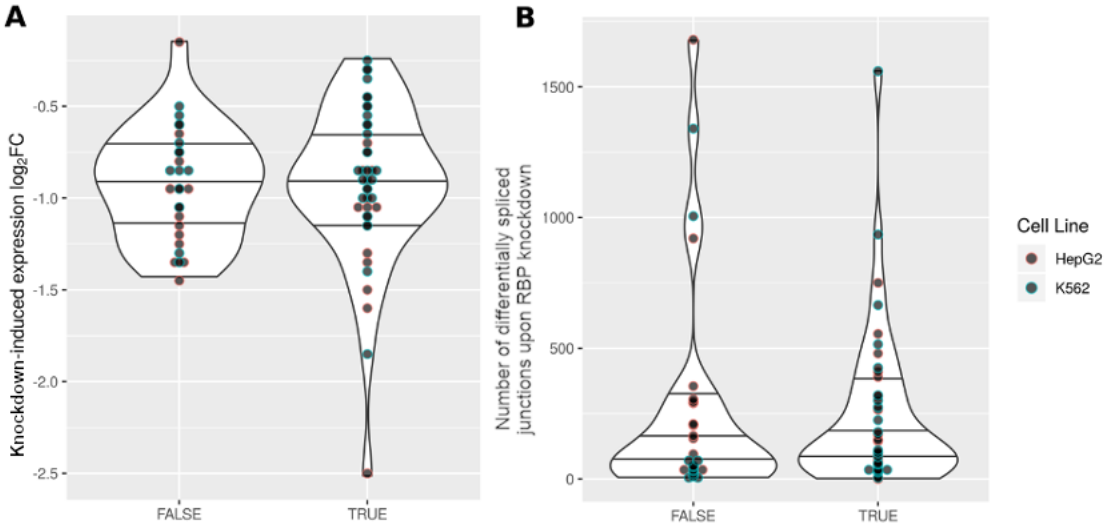

Figure 3-12. Effects of splicing factor knockdown in cases with altered or non-altered motif activity. (A) Knockdown efficiency in splicing factor knockdowns resulting in significant or non-significant change in motif activity. **(B)** Effect of the splicing factor knockdown on splicing in cases of significant vs non-significant changes in motif activity.

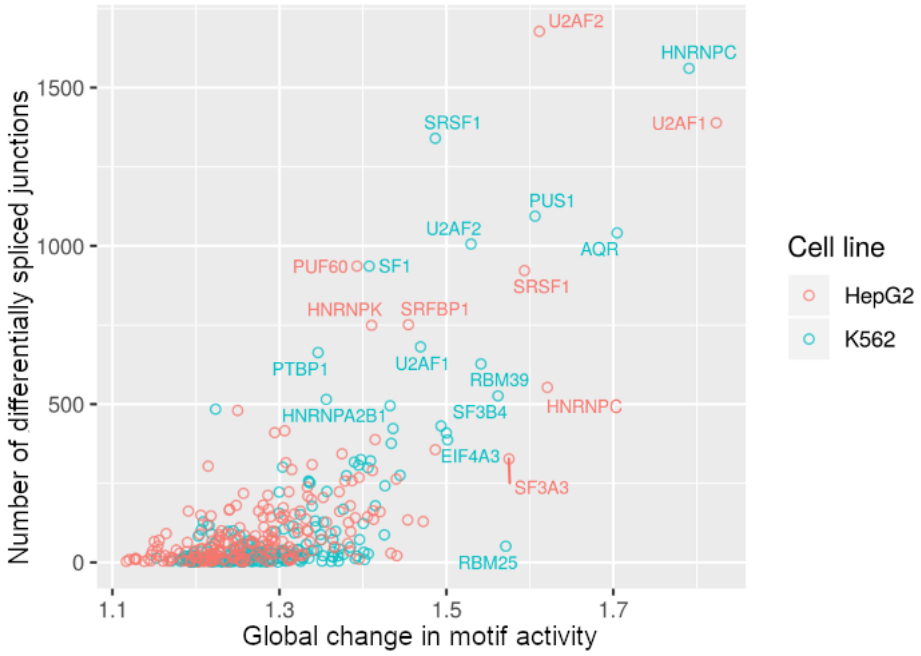

Figure 3-13. Relationship between splicing factor knockdown effect on splicing and motif activities globally. Splicing factors producing greater than 500 differentially spliced junctions or a global change in motif activity greater than 1.5 are labelled.

3.3.3.2 Prediction of splicing factor target splice junctions

Potential target splice junctions of each splicing factor were predicted via a "leave-one-out" analysis (see Chapter 2 for details). The median number of predicted target junctions per motif was ~25,000. The top candidate junctions for each motif, defined as those with target scores > 4 SDs from the mean, were selected for further analysis. This resulted in a reduced, "highest confidence", target splice junction set per splicing factor, with a median of 158 predicted target junctions per splicing factor motif. As expected, predicted target junctions had higher counts of the regulatory motif in question relative to non-target, non-zero count junctions; with a median of three greater motif counts in the predicted target junctions relative to background junctions. The activity of these predicted target junctions upon knockdown of the associated splicing factors was investigated. This showed that predicted target junctions did not have greater absolute changes in PSI values compared to non-target junctions upon knockdown of the motif-associated splicing factor for any splicing factor-knockdowns ($p > 0.05$ - Wilcoxon rank sum test). Additionally, there were no significant overlaps between the predicted target splice junctions and true target junctions, defined as those with significantly altered PSI upon splicing factor-knockdown (FDR > 0.05 - hypergeometric test). Thus, the leave-one-out analysis as currently implemented does not appear to accurately predict the splicing targets of specific splicing factors.

3.3.3.3 Assessing effects of technical confounders on Motif Activity Response Analysis (MARA)

Different motifs have different distributions across the transcriptome, and will vary in the mean and variance of their occurrence across genomic features. Further, the methodology used to count such motifs will also influence the distribution of motif counts. Such variance in distribution may influence how amenable different motifs are to study through MARA. Thus, the potential influence of motif count features on estimation of changes in motif activity was investigated. Motif information content did not differ significantly between motifs with or without a significant change activity upon splicing factor knockdown (mean information content is 7.77 bits and 7.5 bits respectively). To investigate the influence of motif count distribution features on the capacity of MARA to detect changes in motif activity, a PCA regression analysis was performed using matrices of splicing factor motif count occurrences as

input. The first PC of variance was strongly associated with motif count features such as the maximum and standard deviation (SD), as could be expected. However, none of the first 20 PCs were associated with whether a motif had significant knockdown-induced change in motif activity (Figure 3-14). This suggests that features of the motif count distribution did not have a major role in whether a given motif had a significant change in corresponding activity upon knockdown of the associated splicing factor.

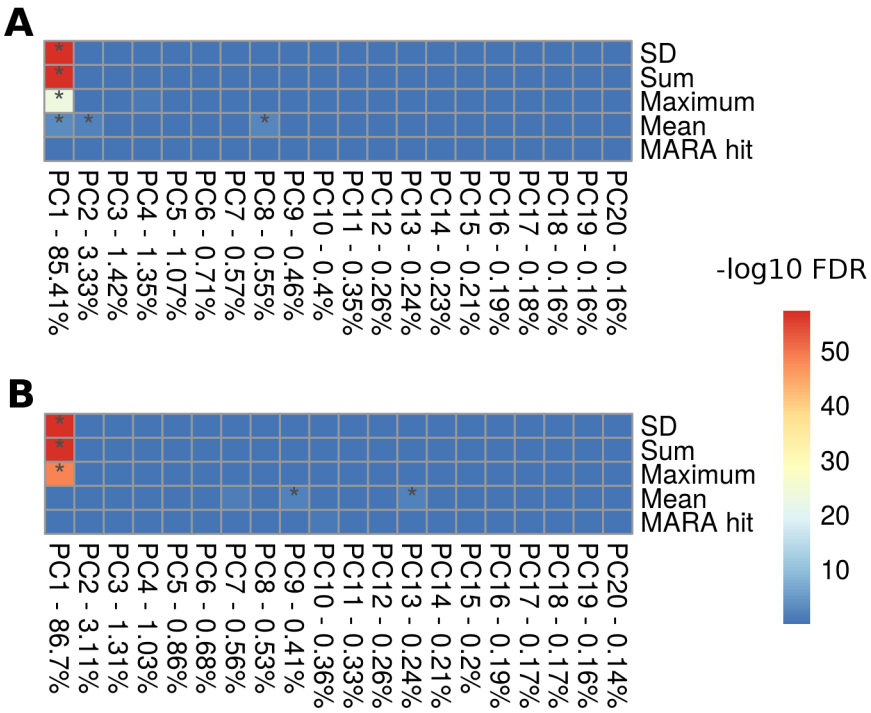

Figure 3-14. PCA regression analysis of motif count features and estimation of changes in motif activities. (A) HepG2 cells, **(B)** K562 cells. Only splice junctions with sufficient read coverage across samples were included in the analysis, and thus count data differs per cell line. Data from linear regression of principal components against count summary statistics (standard deviation [SD], sum, maximum, mean) or whether splicing factor knockdown resulted in a significant change in motif activity ("MARA hit"). * = FDR < 0.05.

Since only two knockdown samples were available per RBP-knockdown, statistical power may be a limiting factor in identifying regulatory splicing factors. A good illustration of these limitations is *HNRNPC*. Knockdown of *HNRNPC* produced similar changes in motif activity in both cell lines (-0.38 in K562, -0.28 in HepG2), but the increased variance in activity amongst knockdown samples in K562 cells appears to have decreased the significance of this difference (Figure 3-15A).

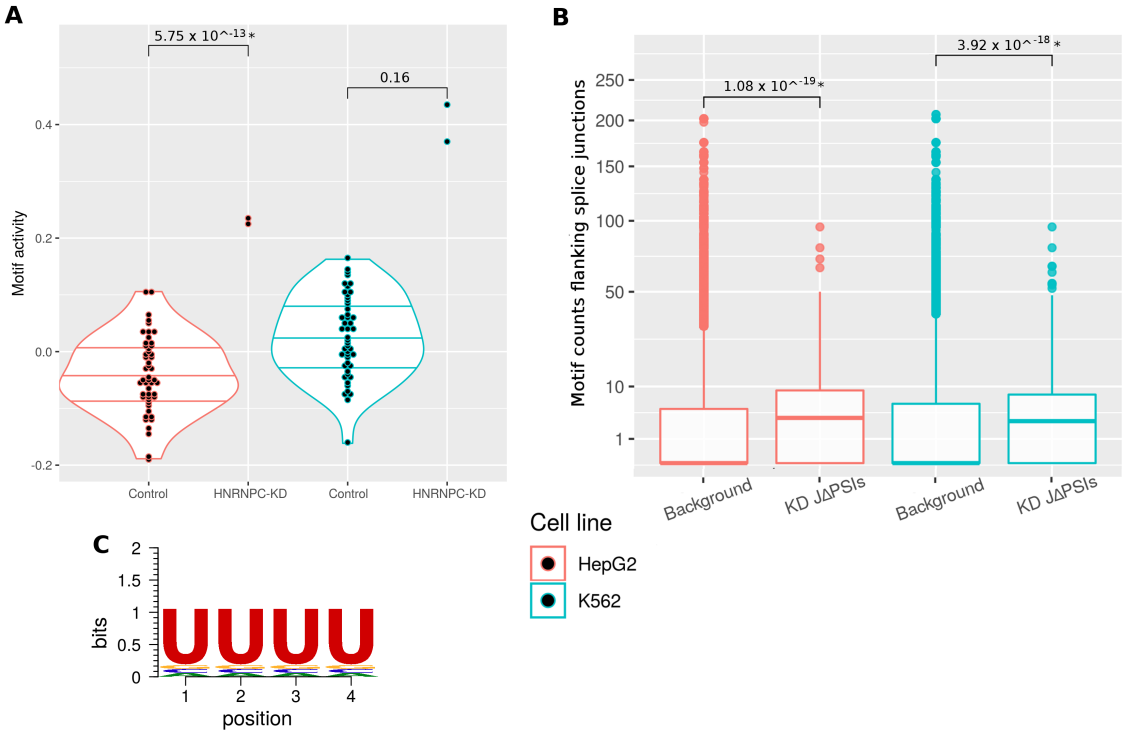

Figure 3-15. Motif-based analysis of *HNRNPC* knockdown. (A) *HNRNPC* motif activity in control and *HNRNPC* knockdown samples. Comparison between groups shows FDR from Student's t-test. **(B)** *HRNPC* motif counts in regions flanking splice junctions with either altered usage upon *HNRNPC* knockdown or unaltered background junctions. Comparison between groups shows FDR of one-tailed Wilcoxon rank sum test. * indicates significant difference. **(C)** *HNRNPC* PSSM motif logo.

3.3.3.4 Motif enrichment analysis of splicing factor knockdowns

To contextualize the results from applying S-MARA, we employed a commonly used motif enrichment procedure. Motif enrichment was used to test for associations between splicing factor-knockdowns and their corresponding motifs. Specifically, the distribution of motif counts was compared between differentially spliced junctions and non-differentially spliced junctions which were used as a background set. Greater motif counts in the RNA regions flanking regulated splice junctions relative to unregulated background splice junctions was taken as evidence of a potential causative role in splicing regulation. For 20 knockdowns in K562 cells and 22 in HepG2 cells, a significant increase in motif counts amongst regulated splice junctions was identified (FDR < 0.05 one-tailed Wilcoxon rank sum test). The overall magnitude of splicing disruption induced by each knockdown influenced the results of this

motif enrichment analysis. Knockdowns with a greater effect were more likely to result in significant cases of increased motif counts amongst differentially spliced junctions (Figure 3-16). This contrasts with the observations made for S-MARA, where such an effect was not found.

In total, 28 different splicing factors had significant enrichment of an associated motif upon knockdown, with 14 cases identified in both cell lines. This is a greater number of successes than that obtained from S-MARA, where 22 splicing factors had significant changes in associated motif activity, with 6 being common to both cell lines (Figure 3-11). This suggests motif enrichment analysis may have greater power to associate specific regulatory splicing factor motifs with differential splicing patterns.

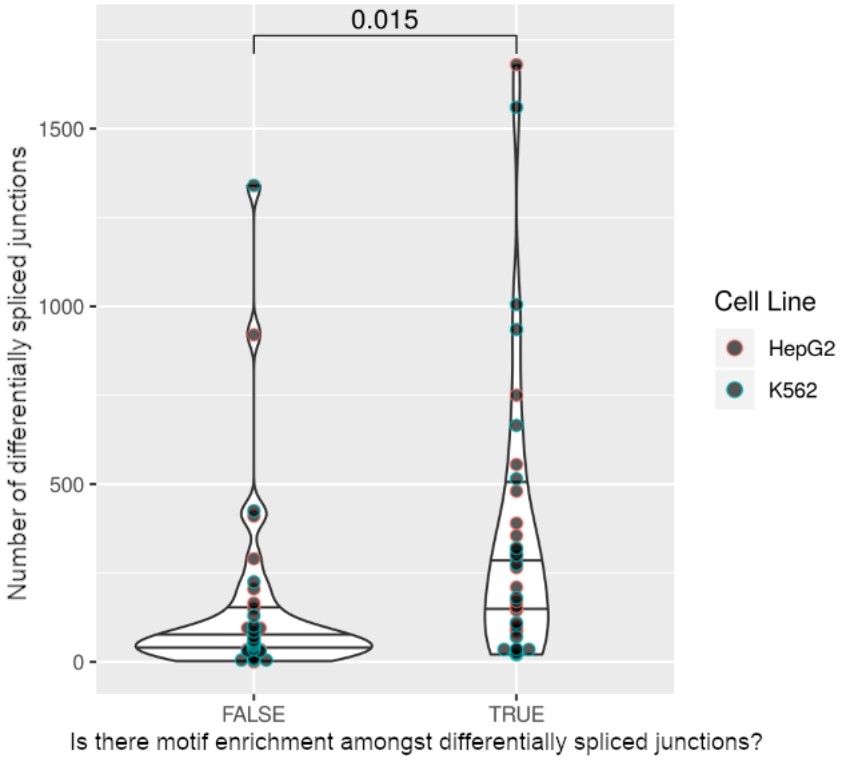

Figure 3-16. Relationship between the effect of each splicing factor knockdown on differential splicing and the results of motif enrichment analysis. Splicing factors are grouped on the x axis according to whether knockdown induced significant enrichment of an associated motif. Wilcoxon rank sum test p value is shown.

We then asked whether motif enrichment and S-MARA identified significant "hits" for the same or different splicing factors. 7/11 MARA-based hits in K562 cells matched the K562 motif enrichment hits, which is a significant overlap as assessed via Fisher's exact test ($p = 1 \times 10^{-4}$) (Figure 3-11). However, for HepG2 cells only 11 of the 17 motif enrichment hits were also identified through S-MARA, which is more consistent with an overlap size expected through chance alone ($p = 0.333$) (Figure 3-11). A number of splicing factor-knockdowns did not result in a significant motif-association using either methodology – with 10 such cases in HepG2 cells and 18 K562 cells. Of interest, *HNRNPC* motifs were enriched amongst differentially spliced junctions in both cell lines (Figure 3-15B), as opposed to the HepG2-specific effect seen with MARA (Figure 3-15A). However, the change in motif activity trended in the same direction in both cell lines for HNRNPC, and the disparity between the two cell lines may relate to the power to detect a significant effect, as discussed above. Interestingly, motif enrichment in differentially spliced junctions was not seen for *HNRNPU* in HepG2 cells (Figure 3-17B) despite MARA detecting a change in motif activity for a *HNRNPU*-associated motif (Figure 3-17A). Activity of the *HNRNPU* motif UGUAUUG showed significant but opposing effects upon knockdown in the two cell lines (Figure 3-17), possibly indicative of opposing splicing enhancer and repressor effects in each cell type.

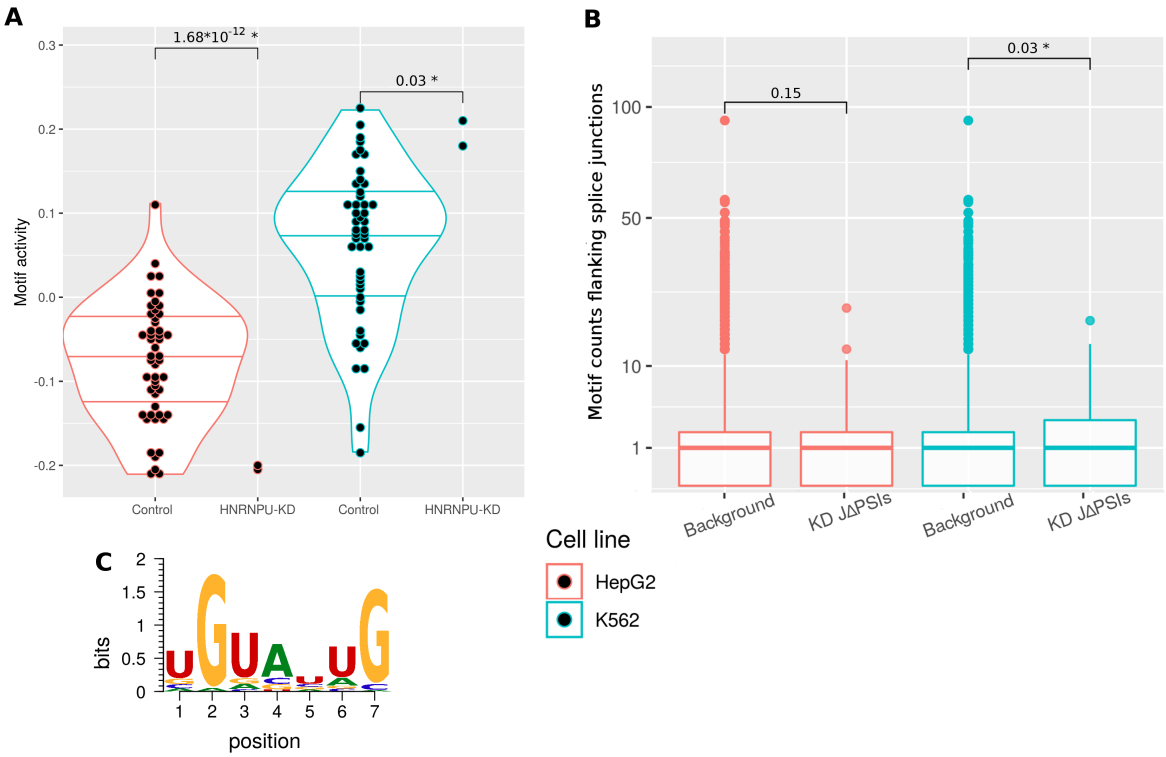

Figure 3-17. Motif-based analysis of *HNRNPU* knockdown. (A) *HNRNPU* motif activity in control and *HNRNPU* knockdown samples. Comparison between groups shows FDR from Student's t-test. **(B)** *HRNPU* motif counts in regions flanking splice junctions with either altered usage upon *HNRNPU* knockdown or unaltered background junctions. Comparison between groups shows FDR of one-tailed Wilcoxon rank sum test. * indicates significant comparison. **(C)** *HNRNPU* PSSM motif logo.

3.3.3.5 Sensitivity and specificity of regulatory motif identification

The motif enrichment analysis approach was able to recover a greater number of regulatory motif "hits" upon splicing factor-knockdown than the MARA-based approach (Figure 3-11). This suggests the enrichment method may have a greater sensitivity to detect regulatory splicing factor motifs. However, the specificity of the two approaches is also of interest. The analysis performed so far was restricted to splicing factor motifs specifically associated with each knockdown-splicing factor. To assess specificity, the effect of each splicing factor-knockdown was assessed across the full set of 103 splicing factor-motifs. Using this approach, those motifs which are specifically associated with each experimentally depleted splicing factor were considered as true positives. These motifs are expected to have a regulatory role in knockdown-induced differential splicing. In contrast, motifs associated with other splicing

factors are not necessarily expected to have roles in regulating such differential splicing, and were thus considered as true negatives.

When analysing all 103 motifs in the context of each knockdown experiment, the motif enrichment analysis identified more significant motifs than the S-MARA approach (Figure 3-18). However, the motif enrichment approach also showed greater receiver operating characteristics (ROC) in identifying those motifs specifically associated with knockdown-splicing factors (Figure 3-19). Indeed, the 95% confidence interval of the area under the ROC curve (AUC) for S-MARA overlapped the 0.5 line. Thus, the MARA approach did not perform better than chance at identifying motifs associated with knockdown-splicing factors (true positives) relative to other splicing factor motifs (true negatives). Using an FDR of 0.1, the true positive rate for detection of knockdown-splicing factor associated motifs was 0.182 and 0.403 for S-MARA and motif enrichment respectively (Table 3-2).

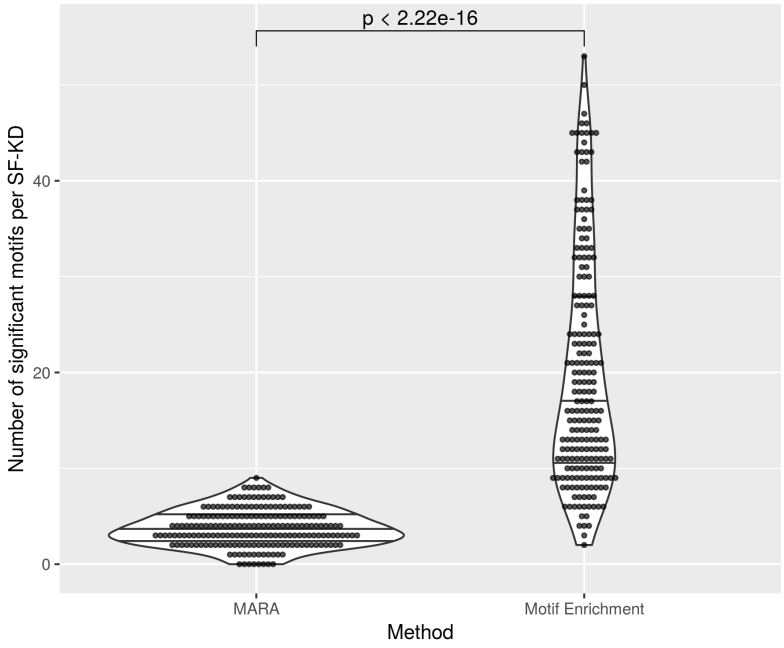

Figure 3-18. Numbers of splicing factor motifs associated with splicing factor knockdowns via S-MARA or motif enrichment analysis. Wilcoxon rank sum test p value shown. SF = splicing factor, KD = knockdown.

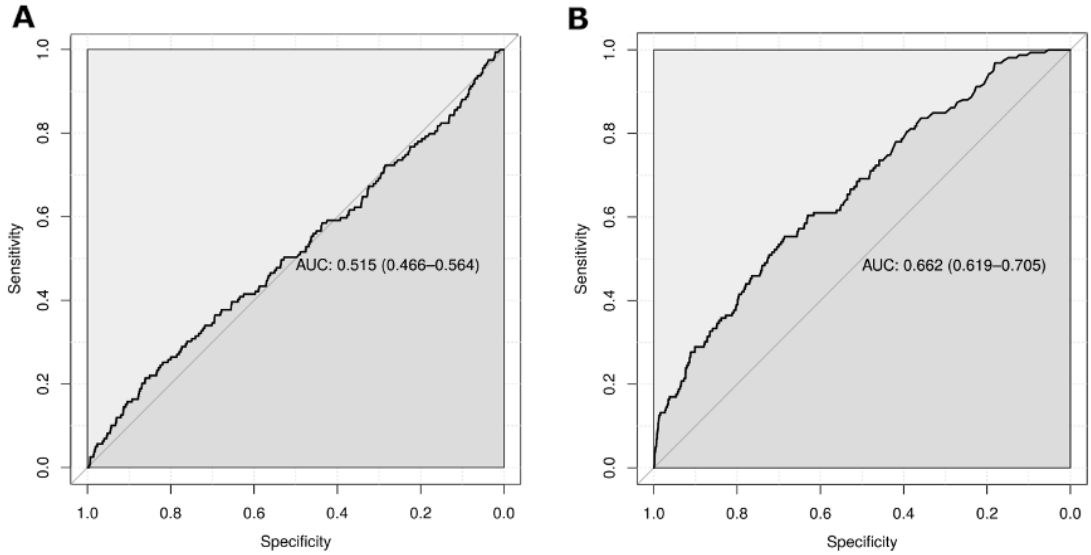

Figure 3-19. Receiver operating characteristics for the identification of regulatory splicing factor motifs. (A) MARA analysis, **(B)** motif enrichment analysis. AUC = area under the curve. 95% confidence intervals are shown. Curves show ratios of sensitivity to specificity at varying FDR threshold values. True positives were defined as motifs associated with each knocked down splicing factor.

Table 3-2. Confusion matrices for the identification of regulatory splicing factor motifs. Numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) splicing factor motifs identified through either MARA of motif enrichment are shown. True positive and true negative rates derived from these values are shown in the bottom rows.

	Method			
	MARA		**Motif enrichment**	
	Predicted regulatory splicing motif	**Predicted not regulatory splicing motif**	**Predicted regulatory splicing motif**	**Predicted not regulatory splicing motif**
Actual regulatory splicing motif	29 (TP)	130 (FN)	64 (TP)	95 (FN)
Not actual regulatory splicing motif	893 (FP)	6261 (TN)	1443 (FP)	5711 (TN)

True positive rate	0.182	0.403
False positive rate	0.125	0.202

Possible reasons for the improved performance of motif enrichment analysis could relate to assumptions underlying the MARA model. For instance, as S-MARA employs linear modelling, there is an assumption of linearity between the PSI of a splice event and the occurrence of splicing factor motifs flanking that splice event. Motif activity estimates are thus coefficients which describe the linear association between motif counts and PSI. This assumption may have several limitations. For instance, the probability of an RBP binding a specific pre-mRNA as a function of the number of motifs it contains will likely be characterised by saturation effects. That is, binding probability will not increase indefinitely with the number of binding motifs. In contrast, the motif enrichment procedure used here does not rely on any assumptions of linearity, being instead based on assessment of rank order through the Wilcoxon rank sum test.

As a simple method of accounting for such saturation effects, the input splicing factor motif count matrix was modified such that the maximum motif count per splice junction region did not exceed a maximum value. Two values for this maximum motif count threshold were trialled: 15 and 30, such that counts were simply capped at the corresponding value. The full workflow was then performed again separately for both maximum values and ROCs re-calculated (Figure 3-20A-B). This modification did not produce a performance increase in MARA.

Another possible advantage of the motif enrichment strategy over MARA is the use of a reduced input set of “positive” splice junctions. This positive set consists of the splice events found to have a significantly altered PSI in knockdown relative to control samples, and is contrasted with a negative set of all other splice junctions. With MARA, in contrast, the motif activity is modelled as the relationship between motif occurrence and splice junction usage for all splice junctions genome-wide. In order to potentially improve the signal-to-noise ratio in the MARA input matrices, a pre-filtering step was performed to include only those splice junctions identified as having significantly altered usage in knockdown samples. This was

performed on a per-splicing-factor knockdown basis per cell line, for a total of 71 analyses (2 cell lines and a mean of 35.5 splicing factor knockdowns per cell line). To clarify, this approach results in the same input splice junctions being used as per the motif enrichment procedure. This per knockdown, "high signal-to-noise", analysis did not produce an improvement in performance of S-MARA (Figure 3-20C).

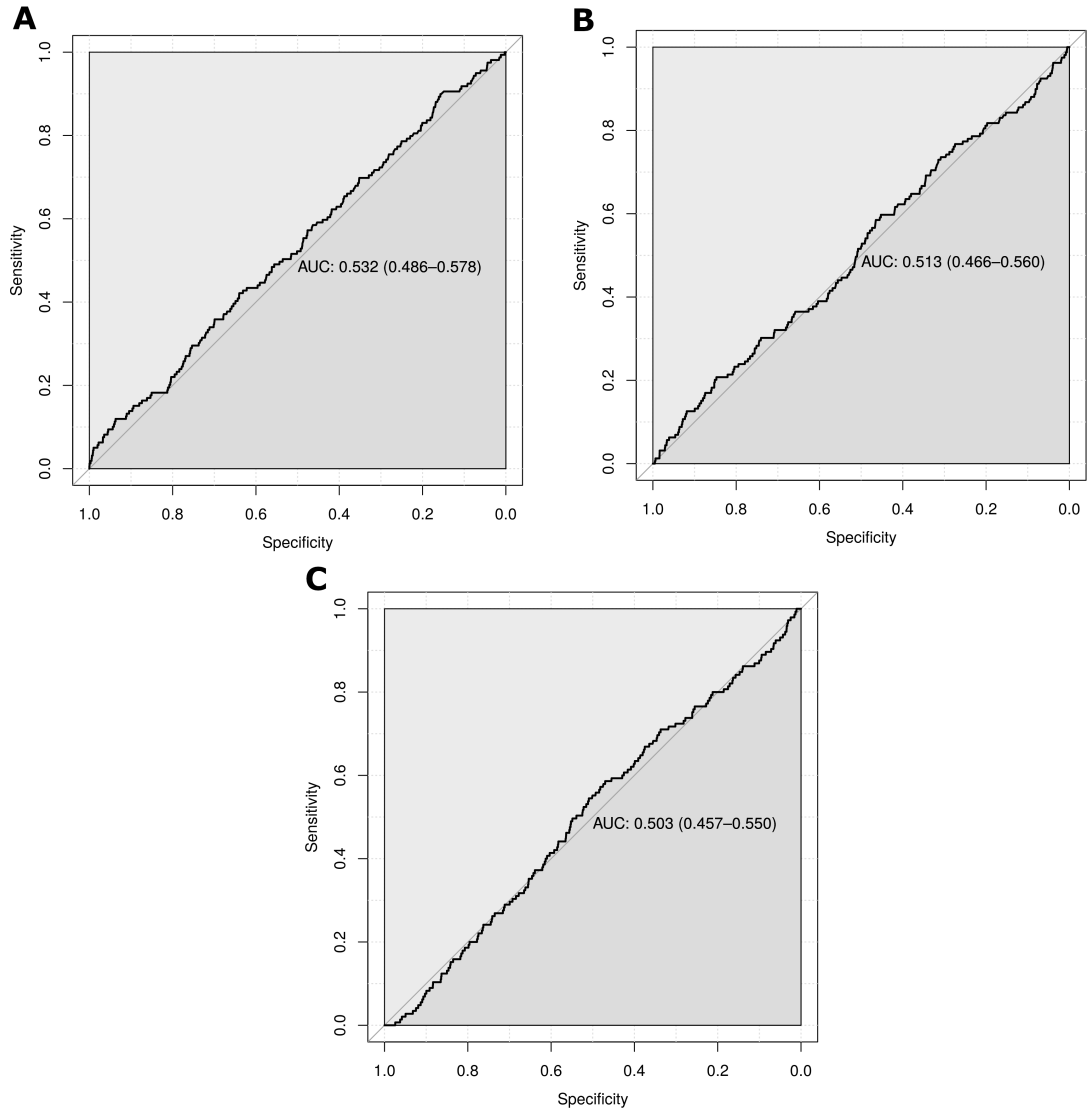

Figure 3-20. Receiver operating characteristics of S-MARA after adaptations to input data. The S-MARA workflow was applied with several modifications made to the input data in an attempt to improve performance. Performance characteristics after setting the maximum number of splicing factor motif matches per-splice site region, to **(A)** 15, or **(B)** 30. **(C)** Performance characteristics after running S-MARA separately for each splicing factor

knockdown, using reduced input data consisting of only those splice junctions with significantly altered use in knockdown samples.

3.3.4 Discussion

In order to apply MARA for the inference of splicing factor motif activity, a pipeline was implemented which involved quantification of genome-wide splicing events, and counting splicing factor motif occurrences in RNA regions flanking these splice events. To facilitate this, a set of 103 splicing factor motifs was compiled. shRNA-induced RBP-knockdown data, generated through the ENCODE project, provided an opportunity to assess the effectiveness of MARA for splicing focused analyses. As expected, reduction in expression of genes encoding RBPs resulted in disrupted RNA splicing, an effect that was more pronounced in RBPs with established roles as splicing factors (Figure 3-5). The magnitude of the effect on splicing was not related to efficiency of the knockdown. However, the overall efficiency of knockdown was not particularly high, with the average reduction in gene expression being less than 50% relative to control samples. Modelling splicing factor motif activity through S-MARA provides a potential method for linking knockdown-induced splicing changes to the regulatory motifs associated with that splicing factor. For a subset of knockdowns, the S-MARA workflow provided such results. That is, motif activities relating to specific splicing factors that had been knocked down were significantly different between control and knockdown samples. A number of technical factors were investigated as potential confounders or sources of limitation. However, knockdown-efficiency, the magnitude of effect of knockdown on splicing, and motif count distribution features, did not affect whether a given splicing factor-knockdown resulted in a significant change in motif activity between control and knockdown samples.

3.3.5 Motif enrichment analysis outperforms splicing-MARA

A motif enrichment approach, based on testing for over-representation of motif occurrences amongst regulated splice junctions, was used as a baseline against which to compare S-MARA. This motif enrichment approach identified a greater number of successful “hits” than the MARA-based approach (Figure 3-11). Further, an analysis of ROC characteristics revealed the motif enrichment procedure displayed greater sensitivity and specificity than S-MARA. Indeed, the MARA-based approach did not perform better than random chance in specifically

identifying splicing factor-knockdown associated motifs (Figure 3-19). Thus, motif enrichment analysis clearly performs better than S-MARA as applied herein.

3.3.6 Limitations in defining true and false positive effects of splicing factor knockdowns

Despite the clarity of the ROC analysis in highlighting the differential performance characteristics of S-MARA and motif enrichment, several limitations to this approach should be considered. In particular, defining true positive and true negative splicing factor motifs has limitations for several reasons. Firstly, splicing factors are commonly under cross-regulation through the activity of other splicing factors in a given gene expression network (Jangi and Sharp, 2014). It is therefore difficult to define true negatives, since such downstream effects may show up as changes in the activity of splicing factor motifs outside of those which were knocked down. Whilst changes in gene expression of off-target splicing factors upon knockdown could be used as a proxy for altered activity, this is imperfect, as altered mRNA abundance does not necessarily indicate a change in splicing factor activity, and conversely, splicing factor activity can be regulated post-transcriptionally. Indeed, detection of these downstream off-target, “false positive”, effects in motif enrichment analysis of splicing factor-knockdown data has been previously observed (Carazo et al., 2018).

Similarly, defining true positives also has limitations. Knockdowns with lower efficiency or with limited effects on splicing would not necessarily be expected to result in a detectable signal through MARA or motif enrichment testing. Indeed, the number of differentially spliced junctions induced by splicing factor knockdown was greater in cases where significant motif enrichment was detected (Figure 3-16). Whilst a similar observation was not found for the relationship between numbers of differentially spliced junctions and change in MARA-estimated motif activity (Figure 3-12B), this may be due to the reduced accuracy of MARA as determined through ROC analysis. With that said, some knockdowns producing low numbers of differentially spliced junctions did result in a significant change in motif activity or motif enrichment, thus highlighting the limitations in using the numbers of differentially spliced junctions as a filter for specifying true positives.

A further challenge with defining true positives is the possibility that a splicing factor may not be acting through binding to the motif associated with that splicing factor in our motif set. For example, FUS, a splicing factor not identified as a hit by either motif analysis method, has the associated motifs CGCGC and GGGGG. However, *in vivo* in the context of neuronal development, FUS was found to bind RNA in a relatively non-specific manner, with only a limited sequence preference for G-rich regions (Rogelj et al., 2012). Therefore, the linear relationship between FUS binding and motif occurrence *in vivo* may be weak, precluding analysis with MARA or motif enrichment analysis. Similarly, many splicing factors have pleiotropic effects and may not be regulating splicing through directly binding to *in cis* elements in pre-mRNA to influence spliceosomal action, but through the regulation of splicing components via other RNA processing mechanisms.

Despite these limitations, the ROC AUC for the motif enrichment analysis (0.662, Figure 3-19B) indicates that there is value in this validation approach. The poor ROC AUC of the S-MARA approach draws into question the reliability of the motif activity estimates. A number of the splicing factor-knockdown MARA hits were specific to the MARA approach and not identified through the motif enrichment method (Figure 3-11) (e.g. *HNRNPU* in HepG2 cells [Figure 3-17]). Given the poor ROC AUC of MARA, it may be that such cases represent false positives.

3.3.7 Possible limitations to the S-MARA methodology

Several possible advantages of the motif enrichment method over S-MARA were investigated. Firstly, since MARA is based upon linear regression, there is an implicit assumption of linearity between splicing factor motif counts and splicing factor activity towards a splice event. This assumption will likely not hold true in many cases. For instance, binding saturation effects will occur, putting a limit to the increase in binding likelihood with increasing motif counts. This potential saturation effect limitation was investigated by placing restrictions on the values of the input motif count matrices. However, this failed to alter the performance characteristics of S-MARA (Figure 3-20A-B). Another possible advantage of motif enrichment analysis is the initial splitting of splice junctions into differentially regulated and non-differentially regulated groups. This initial process could improve the signal to noise ratio. In light of this, a modified S-MARA method was trialled in which input junctions were limited to only those significantly differentially regulated per splicing factor knockdown. With this method, the same input splice

junctions were used as for the motif enrichment procedure, making the two methods more comparable. However, this modification also failed to yield an improved ROC AUC (Figure 3-20C).

3.3.8 Small sample numbers limit statistical power

The ENCODE shRNA experiments were performed in replicates of two. This meant that, whilst control shRNA samples were pooled for analysis herein resulting in greater replicate numbers, estimation of SF motif activity in knockdowns was limited to just two samples. This is of course far from ideal for the estimation of variance as required for statistical analyses such as the Students t-test applied here. As such, issues of statistical power appeared to be a limiting factor in detecting changes in motif activity for a subset of splicing factors. For instance, significant enrichment of *HNRNPC* associated motif counts amongst knockdown-induced differentially spliced junctions was identified in both cell lines (Figure 3-15). Conversely, sample variability appeared to impair the power to detect a significant change in motif activity of these motifs in K562 cells through S-MARA (Figure 3-15). The motif enrichment method uses a comparison of thousands of splice junctions to assess differences in motif counts. In contrast, the applied MARA-based analysis relies upon comparison of two knockdown samples with the group of control samples. It is possible that these methodological differences could make the motif enrichment methodology less sensitive to the low number of knockdown samples available. Future application of these methodologies to a dataset with greater sample sizes is therefore desirable in future.

3.3.9 S-MARA target analysis

Splicing factor-motif target splice junctions were predicted with a leave-one-out-analysis as implemented in the IMAGE pipeline (Madsen et al., 2018). The leave-one-out procedure has been previously validated for prediction of transcription factor targets (Balwierz et al., 2014; Madsen et al., 2018). However, whilst the predicted target junctions were enriched for the specified regulatory motifs in question, they were not enriched for splicing factor targets, defined as those with knockdown-induced differential splicing. The validity of these predicted splice junction targets is therefore questionable. Whilst the target prediction analysis was not the primary focus of this study, the performance of the leave-one-out approach is dependent

upon changes in estimated motif activity. Therefore, any limitations to the motif estimation function of MARA will also affect the target prediction function.

3.3.10 Splicing factor motif enrichment analysis is an effective means to infer regulatory motifs

Whilst the focus of this study was on the development and benchmarking of the S-MARA workflow, the assessment of splicing factor motif enrichment analysis conducted here has its own merit. To our knowledge, the specificity and sensitivity characteristics of a splicing-focused motif enrichment procedure have not previously been investigated. Indeed, this analysis relied upon the use of an extensive resource of splicing factor knockdowns, such as has been provided only recently through the ENCODE project (Nostrand et al., 2018). This work validates the concept of using RNA motifs to infer regulatory splicing factors through analysis of RNA-seq data (Figure 3-19). The motif enrichment approach used here could be further modified in future work to potentially improve upon the sensitivity and specificity characteristics identified here.

3.3.11 Conclusions

Motif activity analysis of splicing factor motifs represents a novel bioinformatic approach to investigating splicing using high-throughput data. The utility of this approach was investigated using a large scale RBP knockdown study. In select cases, S-MARA allowed identification of changes in splicing factor motif activity associated with changes in splicing factor gene expression. However, overall the approach showed poor sensitivity and specificity towards this end. In contrast, a simpler and commonly applied motif enrichment procedure showed reasonable performance in identifying regulatory motifs associated with splicing factor knockdowns. The results of this analysis provide a proof of principle for the utility of splicing factor motif enrichment approaches. A potential limitation of the current analysis relates to the limited sample sizes in the knockdown condition. Despite the identified shortcomings of S-MARA, a further investigation of the approach applied to an experimental system with more biological replicates warrants further investigation

Chapter 4. Motif Activity Response Analysis (MARA) of Splicing Regulators during CD4+ T cell Activation and T_{h2} Polarisation

4.1 Introduction

Activation of CD4+ T cells upon antigenic stimulation of the TCR is characterised by widespread regulation of alternative splicing (Ip et al., 2007; Martinez et al., 2012). The detailed function and control of splicing at several key loci has been well described, with CD45 being the prime example (Hermiston et al., 2003; Lemaire et al., 1999; Oberdoerffer et al., 2008; Rothrock et al., 2005; Wang et al., 2001). Likewise, polarisation of CD4+ T cells into functional subsets is associated with broad splicing regulation (Stubbington et al., 2015). Further elucidation of the regulatory splicing factors underlying control of such splicing regulation is necessary to fully understand the gene expression programmes which drive and maintain CD4+ T cell states.

Henriksson *et al.* (Henriksson et al., 2019) recently performed a detailed timecourse investigation of the transcriptional regulatory programme driving both CD4+ T cell activation and polarisation to the T_{h2} lineage. A comprehensive sampling of the activation and polarisation timecourse was performed in primary human CD4+ T cells from three donors, with 10 time points post-activation (0, 0.5, 1, 2, 4, 6, 12, 24, 48, and 72 hrs) being profiled through RNA-seq. Cells were activated through direct CD3 and CD28 stimulation, followed by IL-2 treatment 48 hrs later. A subset of cells was additionally treated with IL-4 at time point 0 to initiate T_{h2} polarisation. These data provide a rich resource to study other aspects of the gene expression pathway, including splicing.

Although initial benchmarking detailed in Chapter 3 drew into question the accuracy of the S-MARA methodology, a potential limiting factor identified was the low number of biological replicates per experiment. The Henriksson *et al.* timecourse has greater replicate numbers (three) and more biological conditions (10 time points in two cell states). These features may make this a more powerful data set through which to test S-MARA. Further, the efficacy of splicing factor motif enrichment was demonstrated through analysis of splicing factor knockdowns, as presented in Chapter 3. Therefore, both of these motif-based methods will be applied to analysis of the Henriksson *et al.* timecourse.

The program of gene expression following CD4+ T cell activation is dynamic, involving rapid transcriptional regulation to thousands of genes (Henriksson et al., 2019; LaMere et al., 2016, 2016; Ni et al., 2016). Further, the activation process is thought to be modulated by distinct regulatory feedback processes acting through key protein nodes (Martínez-Méndez et al., 2020). We therefore hypothesise that a network analysis approach aimed at identifying modules of co-regulated splicing events will reveal distinct groups of events regulated with different temporal dynamics and via distinct feedback mechanisms. To this end I will apply the network analysis tool Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008; Zhang and Horvath, 2005) to both splice event quantifications and motif activities. WGCNA has been previously applied to identify groups of co-regulated genes across timecourse data (Smith et al., 2016).

4.2 Aims

Hundreds to thousands of genes are documented as being differentially spliced during the T cell activation process, and the mechanisms of control underlying many of these splicing events remains to be fully characterised. Here, I aim to profile the splicing regulatory network across a timecourse of CD4+ T cell activation. To this end I will combine a correlation-based splicing module discovery analysis with splicing factor motif-based analyses. I will utilise both S-MARA and motif enrichment analysis with the aim of identifying both known and novel candidate splicing factor regulators of the CD4+ T cell activation process. Specifically, I aim to:

1. Identify and investigate modules of co-regulated splice events across the CD4+ T cell activation and polarisation timecourse.
2. Identify candidate splicing factor regulators of such co-regulated modules through both S-MARA and motif enrichment analysis.
3. Assess the ability of both S-MARA and motif enrichment analysis to identify known splicing factor regulators of activation-dependent splicing changes in CD4+ T cells.
4. Apply S-MARA and motif enrichment analysis to identify novel candidate regulatory splicing factors.

4.3 Results

4.3.1 Genome-wide splicing profiles during CD4+ T cell activation and polarisation

In order to capture and profile the potentially complex and diverse patterns of splicing regulation in CD4+ T cells undergoing an activation response, a correlation-based network approach was employed. To this end, we made use of WGCNA in order to identify modules of co-regulated splicing events. After initial filtering to remove junctions which were quantifiable in less than 50% of samples, a set of ~108,000 junctions from 9392 genes were used for downstream analysis. Using these junctions, PCA indicated that the main source of variance in splice junction PSI was time after activation (Figure 4-1).

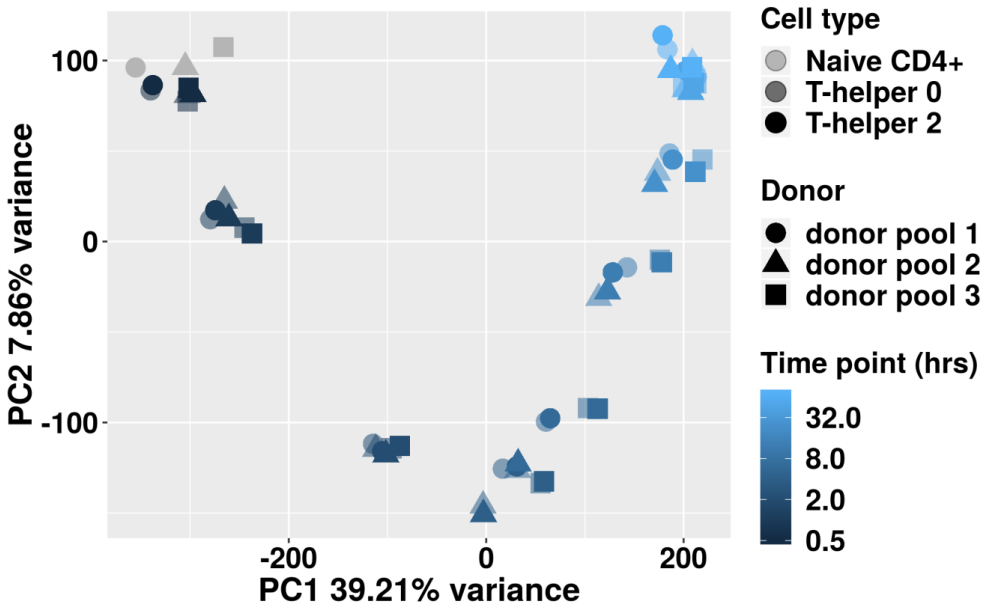

Figure 4-1. PCA of splice junction logit transformed PSI values from primary CD4+ T cells during a timecourse of activation and polarisation. Donor pool = set of 12 technical replicates from an individual donor pooled for analysis after RNA-seq and alignment. Cell type is indicated by opacity of data points.

Application of WGCNA identified 30 splice junction modules — groups of splice events with similar PSI profiles over time after CD4+ T cell activation. Since modules resulting from WGCNA vary in size, a common approach to summarise module behaviour is to use the first prinicipal component of variance of each module. In WGCNA nomenclature, this value is referred to as

the module eigenGene; or here as the eigenJunction. The eigenGene expression is then used as a summary metric for the coordinated behavior of a given module. Initial observation of module eigenJunctions revealed that some of these modules were invariant across the timecourse, and instead showed patterns of differential splicing between biological donor. In order to identify splicing modules with variation over time, linear mixed effect spline modelling was utilised. The use of a spline-based method facilitates modelling of varied and complex temporal relationships, and this approach has been applied to analysis of the activity of varied biological molecules in timecourse experiments (Straube et al., 2015). This linear modelling approach identified 21 modules as having a significant relationship between the eigenJunction, and time after TCR stimulation (FDR < 0.05). As expected, other module eigenJunctions were predominantly correlated with donor of origin. These 21 modules were selected for further analysis. Individual splice junction PSI values were also tested for significant associations with time-after-activation with linear mixed effect spline modelling. Of the remaining 21 modules, 15 had greater than 50% of their individual member splice junctions with a PSI significantly associated with time-after-activation. These 15 modules were selected for further analysis (Figure 4-2). Only a single module had an eigenJunction with a significant cell type vs time interaction effect, and this appeared to be driven by a single T_{h2} sample (Figure 4-2 – module 14, FDR = 0.022). Thus, the splicing regulatory profile of activation appears to be similar in both naïve CD4+ T cells and those undergoing polarization to a T_{h2} subtype.

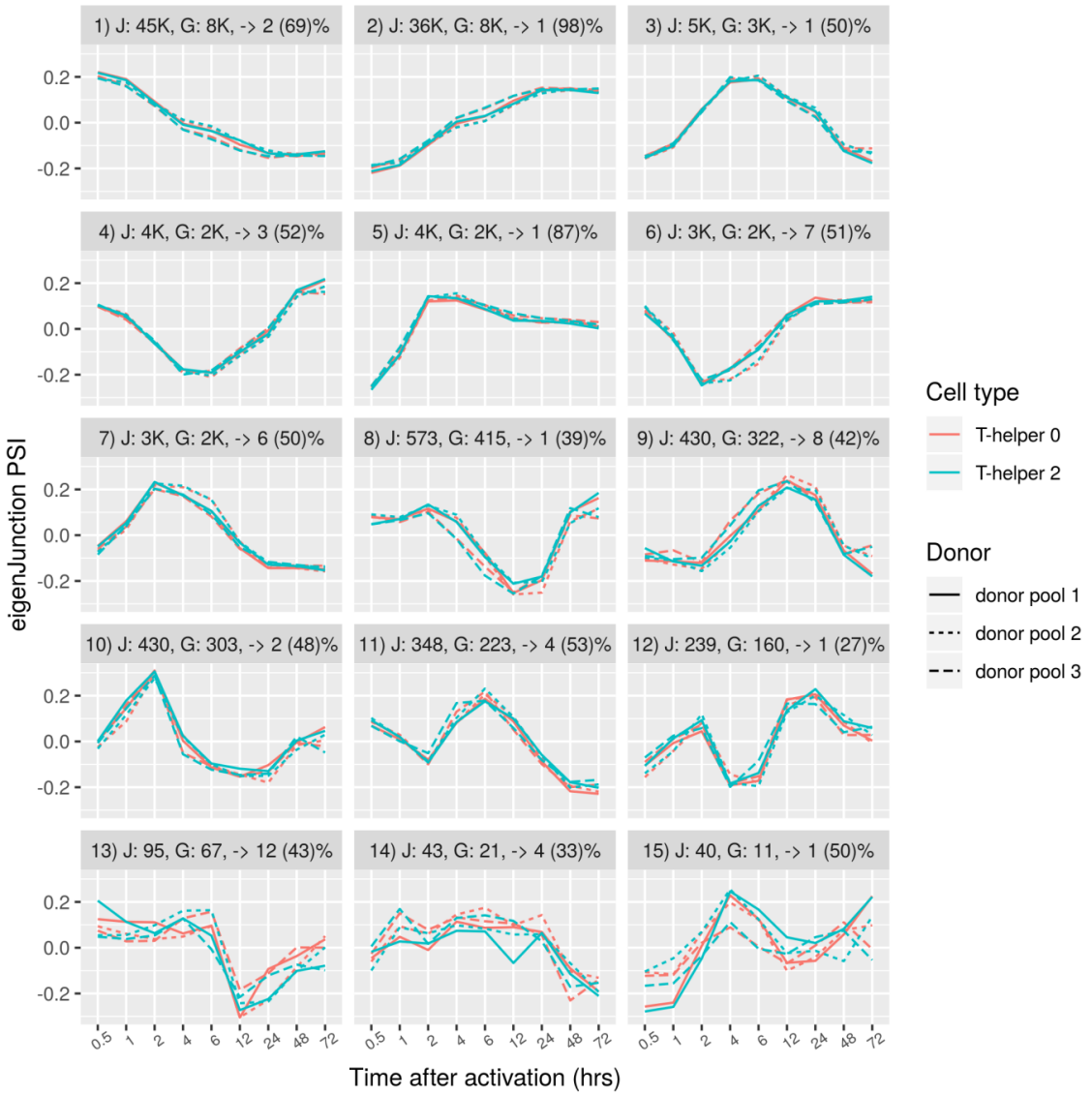

Figure 4-2. Junction splicing profiles during CD4+ T cell activation and polarisation. Splice junctions are grouped into modules via WGCNA, with modules then being represented by the first principal component of variance of member junctions – the "eigenJunction" PSI, which is on a logit scale. Modules have been filtered to show only those with significant relationships with time-after-activation. J = number of splice junctions in each module, rounded to the nearest 1000 (K) for larger modules. G = number of genes represented in each module, rounded to the nearest 1000 (K) for larger modules. -> highlights the module with the highest similarity in terms of percentage of junctions which are part of the same local splicing variation (LSV). This occurs due to cases where individual splice junctions of an LSV are in different

modules. Donor pool = set of 12 technical replicates from an individual donor pooled for analysis after RNA-seq.

These 15 modules (Figure 4-2) collectively detail splicing of 8882 genes, ~50,000 local splicing variants (LSVs), and ~80,000 unique splice junctions (out of the 9392 genes, ~61,000 LSVs, and ~108,000 splice junctions with sufficient read data to be quantifiable). Splice junction modules displayed a range of profiles, including steady changes in splicing behaviour (e.g. modules 1 and 2), transient switches (e.g. modules 3, 4, 6, 7, 8, and 9), or more complex oscillating patterns (e.g. modules 10, 11, 12, and 15) (Figure 4-2). Several junction modules had inversely correlated eigenJunction profiles with one another, such as modules 1 and 2, 3 and 4, 6 and 7, or 8 and 9. This effect is at least partially driven by the structure of LSVs as defined by MAJIQ. The PSI of a junction is by definition dependent upon the PSIs of other junctions from the same LSV. For instance, as the usage of a potential alternative and upstream 5' splice site increases; the relative usage of the corresponding potential downstream 5' splice site necessarily decreases. This effect is most pronounced between modules 1 and 2, with 98% of module 2 junctions being members of LSVs with junctions present in module 1 (Figure 4-2).

Splice junctions were mapped to genes prior to performing a gene ontology analysis. This revealed enrichment for distinct biological processes within splice junction modules (Figure 4-3). "Steady-switch" modules 1 and 2 were enriched for genes involved in many steps of the gene expression pathway, including transcription, splicing, translation, and other co-and-post-translational processing. The "inverted-U" shape module 3 is enriched for genes involved in histone acetylation, "U-shaped" module 4 contains genes relating to co-translational protein transport, and module 5 is also enriched for genes involved in translation as well as splicing, consistent with auto or cross-regulation of splicing components.

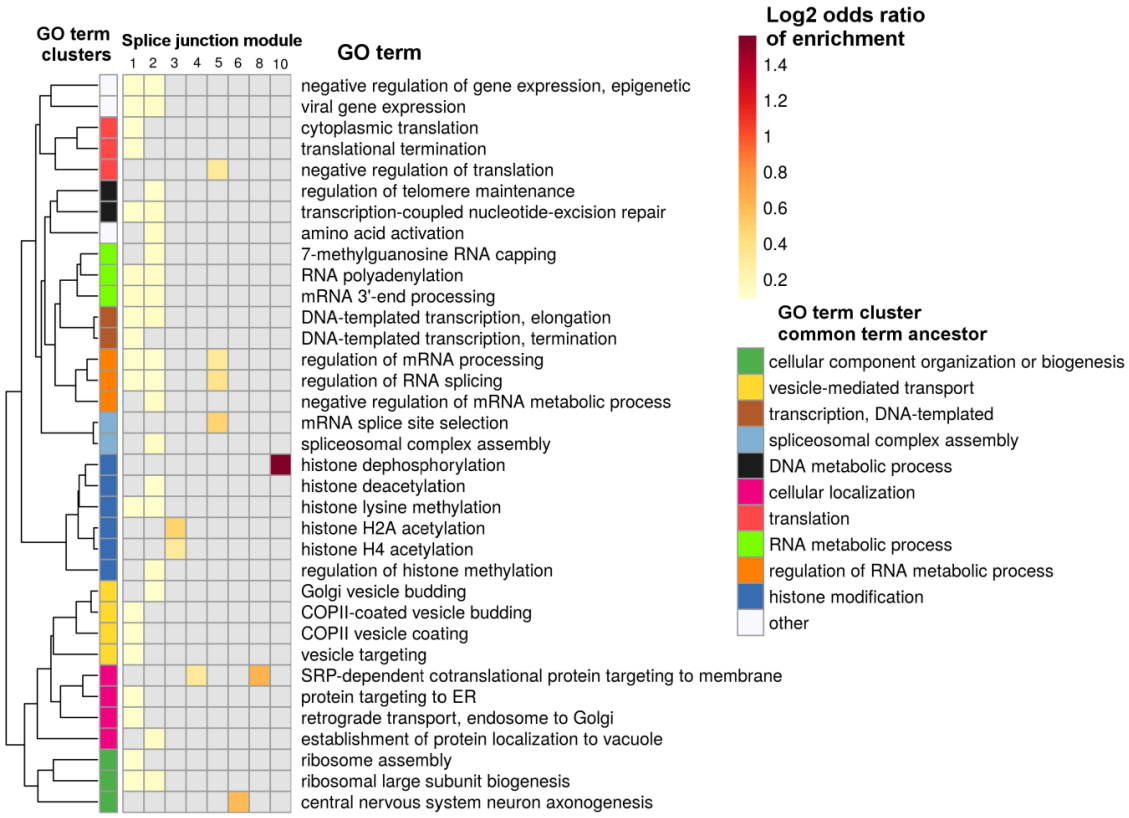

Figure 4-3. Gene ontology enrichment analysis of splice junction module genes. Modules with at least one enriched GO term are shown. Grey boxes indicate lack of significant enrichment for a GO term in a given module. GO terms are clustered via their "semantic similarity" within the GO graph structure. Similar GO terms are labelled with the most specific term that collectively describes that group of terms – the "common term ancestor" in the GO graph structure. Enrichment analysis was performed using the "Biological Processes" ontology.

4.3.2 Splicing factor motif activity profiles during CD4+ T cell activation and polarisation

Splice junction PSI values across the timecourse of activation were used to define splicing factor motif activities through S-MARA. To explore the properties of these motif activities across the timecourse, a similar module-based analysis was performed. Since there are many fewer splicing factor motifs relative to splice junctions, a reduced form of the WGCNA workflow was employed (see Methods). This resulted in 11 modules of motif activity, with a median of 7 motifs per module (range 3-38), and a range of profiles over the timecourse (Figure 4-4). Linear mixed effect spline modelling was performed to test for association of module eigenMotifs (first PC of variance of motif activities) with time after TCR stimulation.

This identified 11 modules as having a significant relationship with time after activation (FDR < 0.05), with module 7 being the sole non-significant case (Figure 4-4). Modules 2, 3, and 5 had a significant cell type versus time interaction (FDR < 0.05). This effect appeared relatively subtle however, and within these modules, time after activation rather than cell type appeared to be the main driver of motif activity (Figure 4-4). Patterns of splicing factor motif activity included steady and gradual changes (e.g. modules 1, 2, 3 and 10), transient changes (e.g. modules 4, 5, 6), and more complex patterns of modulation (Figure 4-4). Of interest, motifs within modules often shared similar sequence content (Figure 4-5). For instance, module 5 contained motifs with prominent GC dinucleotide occurrences (Figure 4-5). Motif module 1 contained motifs of diverse sequence composition (Figure 4-5), whilst module 2 contained more homogenous clusters of motifs such as a group of U/AGGG based motifs associated with members of the HNRNP family (Figure 4-5).

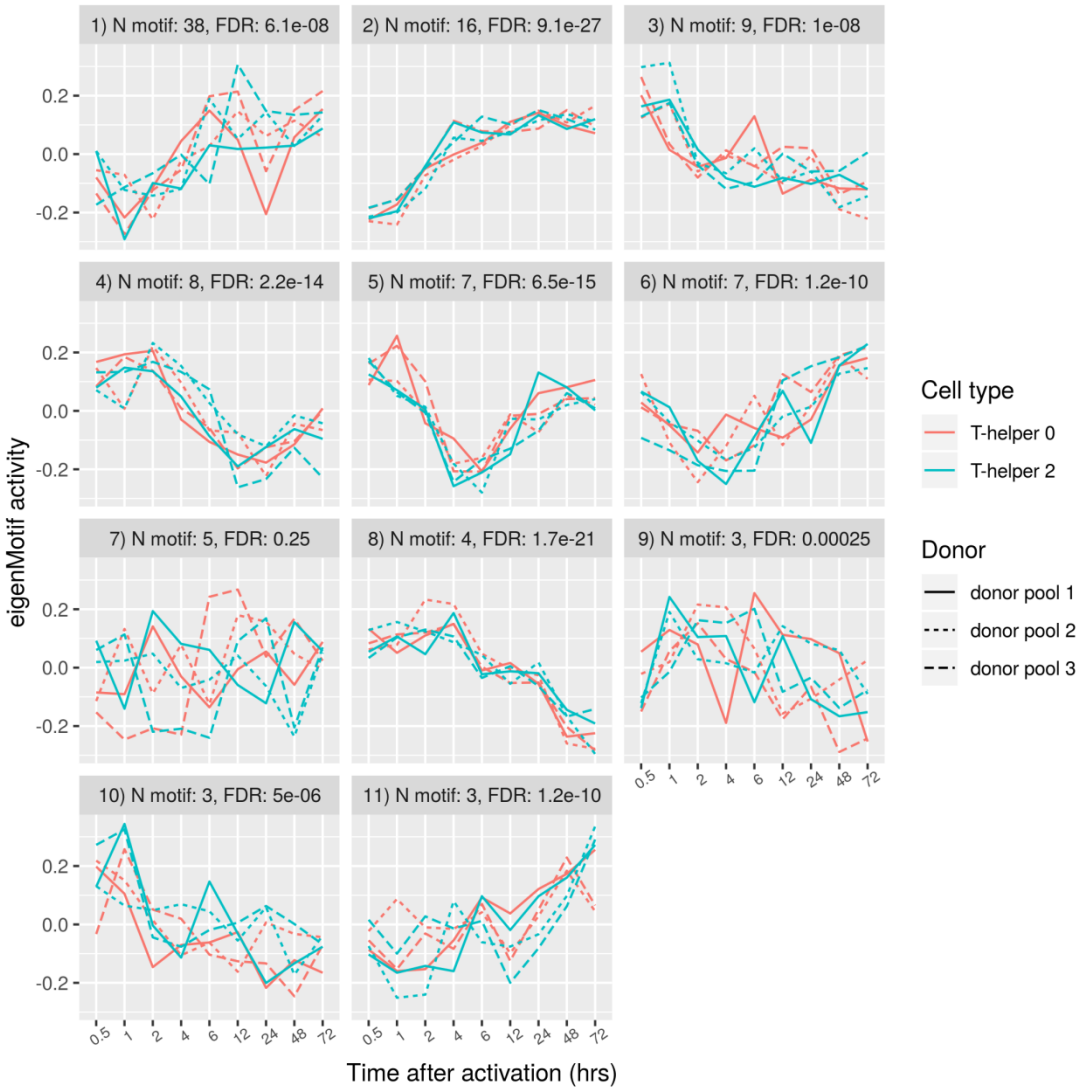

Figure 4-4. Splicing factor motif activity profiles during CD4+ T cell activation. Motifs are grouped into modules via hierarchical clustering, which are then represented by the first principal component of variance ("eigenMotif"). Motif activity values are scaled and centred prior to calculation of module eigen-motifs. N motif = number of splicing factor motifs in each module. FDR = false discovery rates relating to the null hypothesis that eigen-motif activity has no relationship with time after activation - assessed via linear mixed effect spline modelling. FDR < 0.05 except for module 7. Donor pool = set of 12 technical replicates from an individual donor pooled for analysis after RNA-seq.

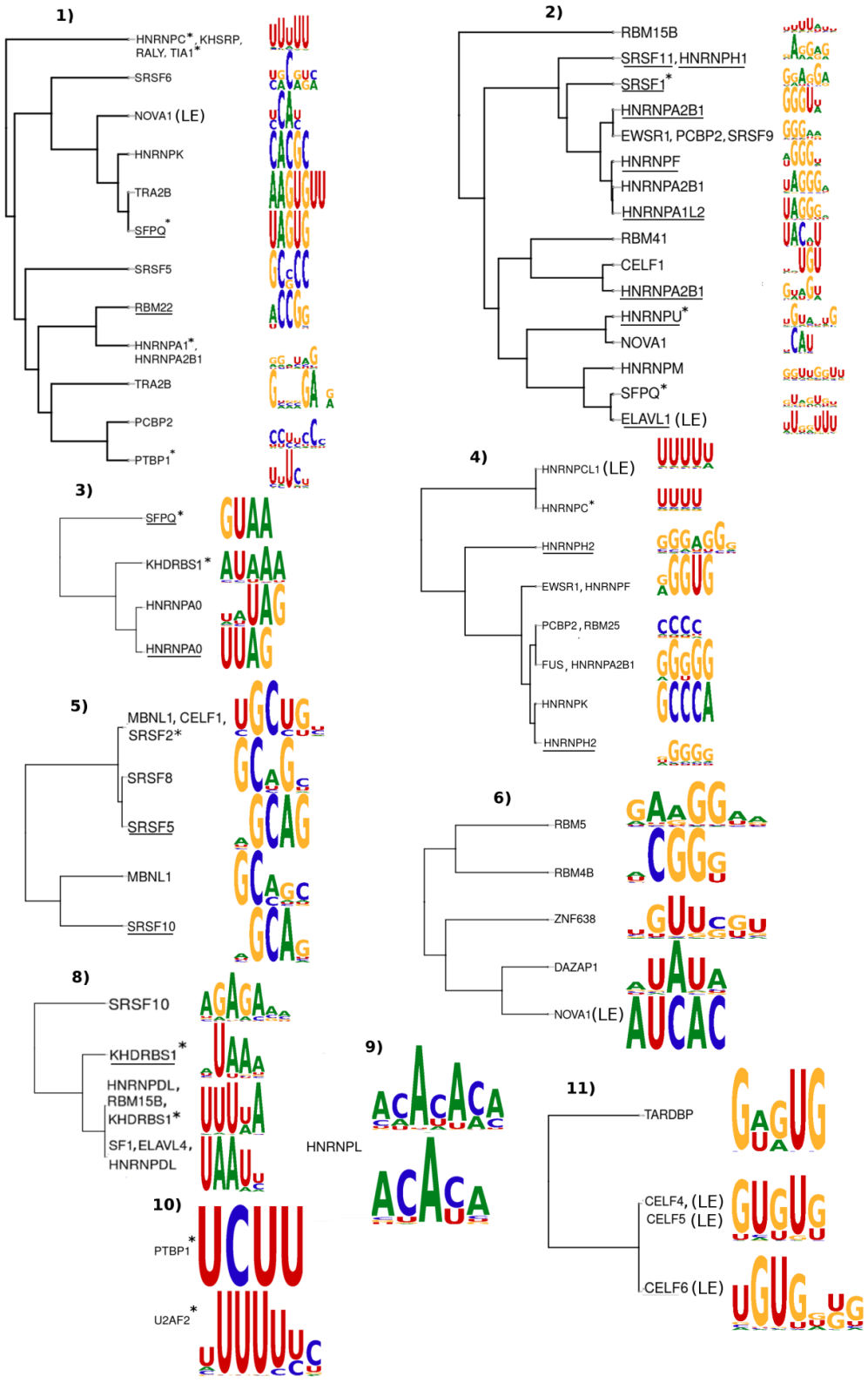

Figure 4-5. Motif logos of splicing factor motif activity modules. Modules as defined in Figure 4-4. Nucleotide height is proportional to information content in bits. To allow visualisation, motifs were filtered to show only those which were most significantly associated with time

after CD4+ T cell activation (defined as motifs with an FDR for association with time after activation < 5 x 10^{-4} as assessed via linear mixed effect spline modelling). Module 7 not shown since this module did not have a significant eigenMotif time profile. Motifs within a module are clustered according to Pearson correlation of the position weight matrices after Smith-Waterman local alignment. Motifs are labelled with the associated splicing factors. To facilitate visualisation, the splicing factors used for labelling are also filtered. Splicing factors which are expressed and have gene expression correlated with motif activity are preferentially used for labelling, unless this resulted in removal of all splicing factors from a given motif. Splicing factors which are unexpressed or with low expression are indicated (LE). Splicing factors for which gene expression is correlated with motif activity are underlined. Pre-defined positive control splicing factors are marked with * (Table 4-1).

4.3.3 Identifying potential regulatory interactions between splicing factor motifs and splice junction modules

4.3.3.1 MARA-based inferences

To gain insight into potential regulatory interactions between splicing factors and modules of splice junction activity during CD4+ T cell activation, a correlation-based approach was used. Correlation between a splicing factor motif activity and splice junction PSI suggests a potential role for the given splicing factor in regulating the splicing of that junction through binding the associated motif. Significant correlations between individual splicing factor motif activity and junction module eigenJunction PSIs were observed (Figure 4-6). Two clusters of junction modules having broadly inverse patterns of correlation with motif activity were present (Figure 4-6). As expected, these two groups of junction modules contained module pairs with anti-correlated PSI profiles such as junction modules 1 and 2. Splice junction modules 1 and 2 stand out as cases in which eigenJunction PSIs are highly correlated with a number of motif activity values, as do modules 3 and 4.

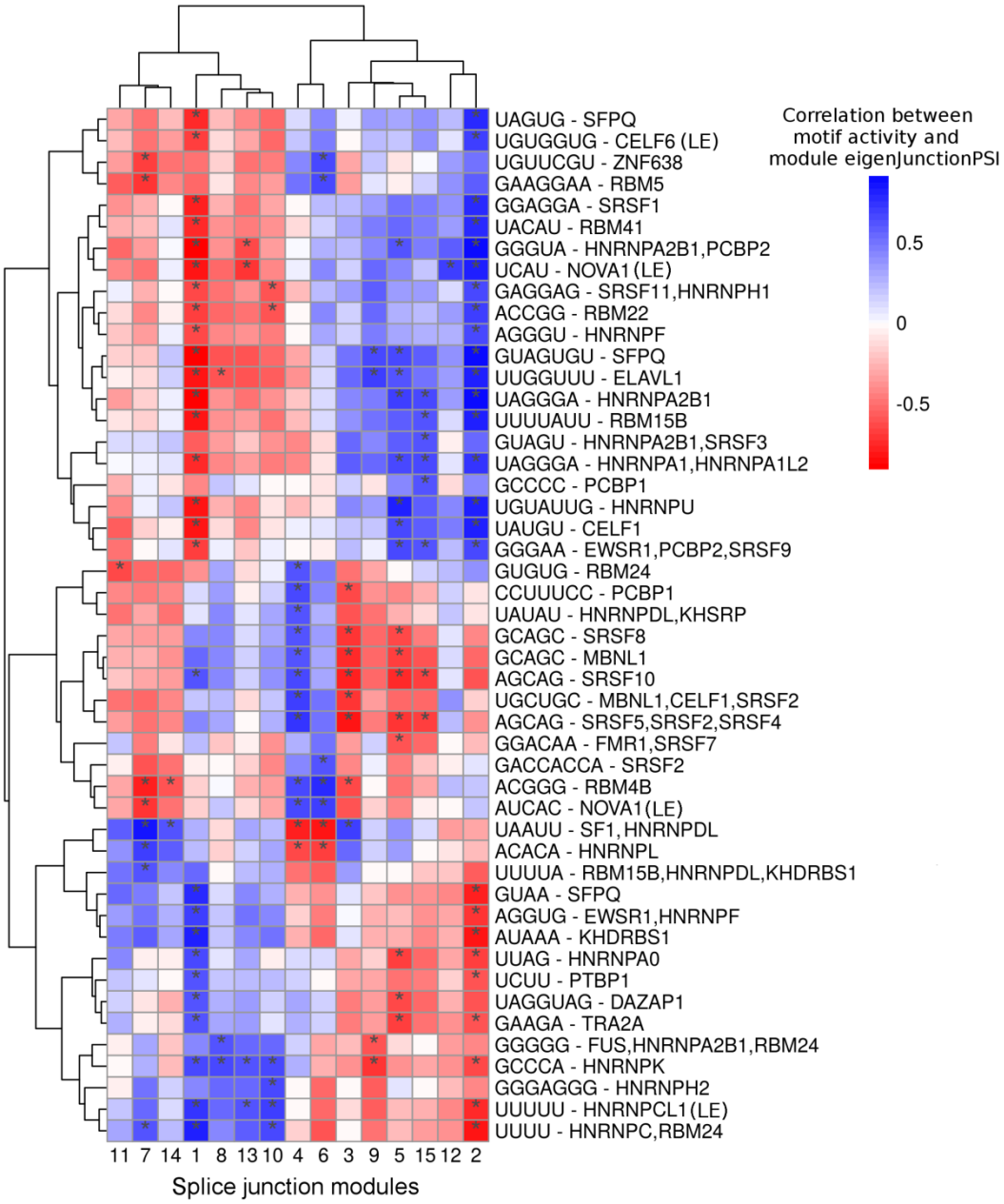

Figure 4-6. Correlation between splice module eigenJunction PSIs and splicing factor motif activities. * indicates an absolute Pearson correlation > 0.6. Consensus motifs and associated splicing factors are shown. To simplify these row labels, and since multiple splicing factor can be associated with a given motif, splicing factors which are expressed are preferentially used for labelling, unless this resulted in removal of all splicing factors from a given motif. Splicing factors which are unexpressed or with low expression are indicated (LE). Only motifs with an absolute Pearson correlation > 0.6 with at least one splice junction module eigenJunction value are shown.

4.3.3.2 Motif enrichment-based inferences

A motif enrichment approach to infer potential regulatory interactions between splice junction modules and splicing factor motifs was also employed. To this end, the distributions of motif occurrences flanking splice junctions within each module were compared against a background distribution consisting of all other non-module junctions. This approach revealed that splice junction modules were enriched for a number of splicing factor motifs in flanking RNA sequences (Figure 4-7). Junction module 1 showed enrichment for a number of motifs, most strongly for motifs composed of U-stretches such as the UUUU motif associated with positive control splicing factor *HNRNPC* (Table 4-1). This is also observed with junction modules 5 and 12, although the magnitude of enrichment is stronger in these modules. The strongest examples of enrichment are seen in the smaller junction module 15, which is strongly enriched for motifs associated with *KHDRBS1* (AUAAA), *DAZAP1* (AUAUA), and *HNRNPD* (UAUUA).

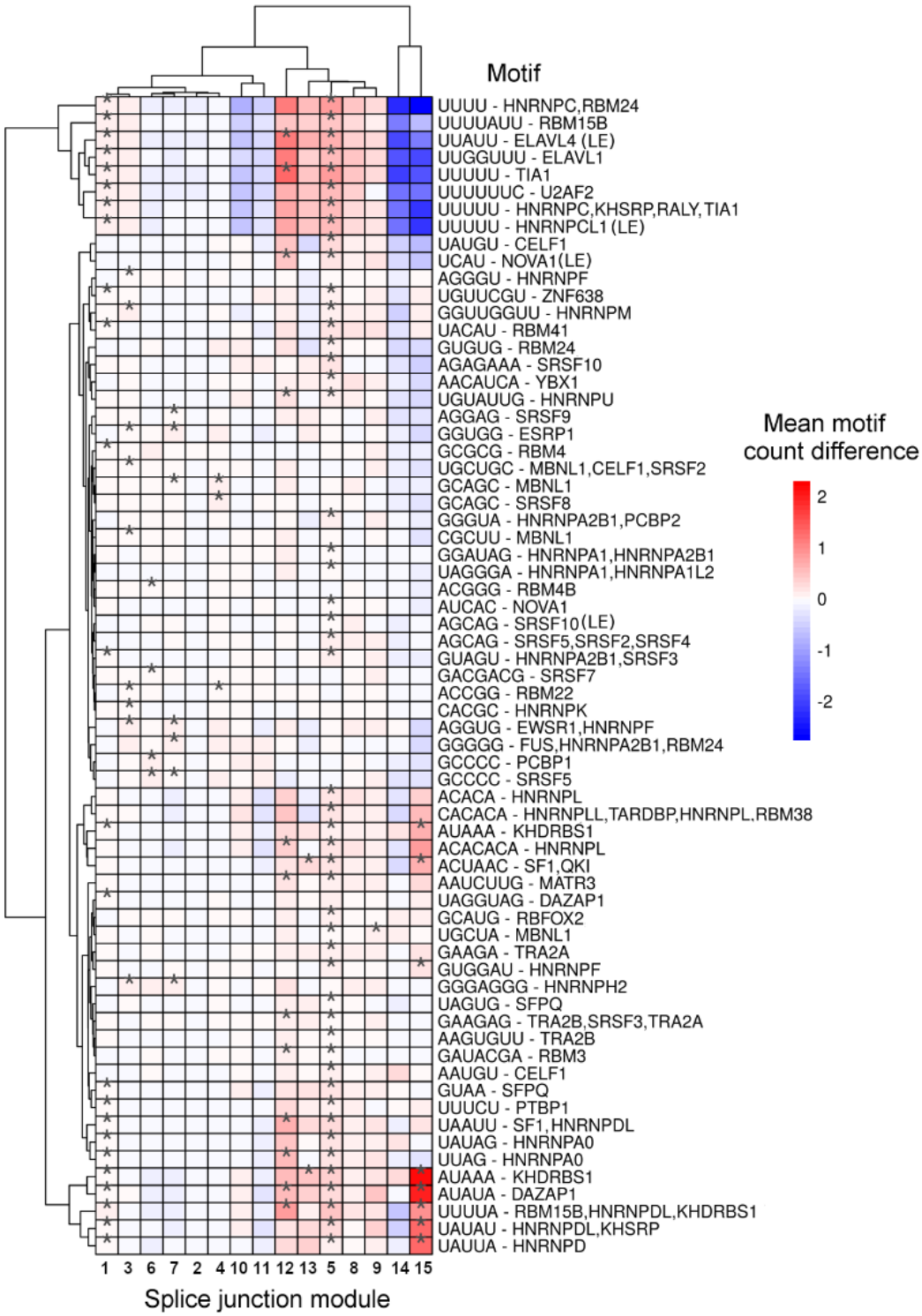

Figure 4-7. Splicing factor motif count enrichment in splice junction modules. Modules having splice junctions with greater motif counts relative to non-module splice junctions are indicated with * (FDR < 0.05, one-tailed Wilcoxon rank sum test). Mean motif count difference refers to count flanking module member splice junctions relative to non-module junctions. Splice

junction modules were also underrepresented for counts of some motifs, as indicated by blue squares, although such underrepresentation was not directly tested for. Consensus motifs and associated splicing factors are shown. To simplify row labelling, and since multiple splicing factor can be associated with a given motif, splicing factors which are expressed are preferentially used for labelling, unless this resulted in removal of all splicing factors from a given motif. Splicing factors which are unexpressed or with low expression are indicated (LE).

4.3.3.3 Comparison of MARA and motif enrichment approaches

Both the MARA-based approach and the motif enrichment-based approach provide methods for associating motifs and splicing factors to groups of splice junction modules for inference of potential regulatory relationships. Results from the two approaches associated different numbers of splicing factor motifs with each splice junction module. The two methods identified largely independent sets of potential regulatory motifs (Figure 4-8).

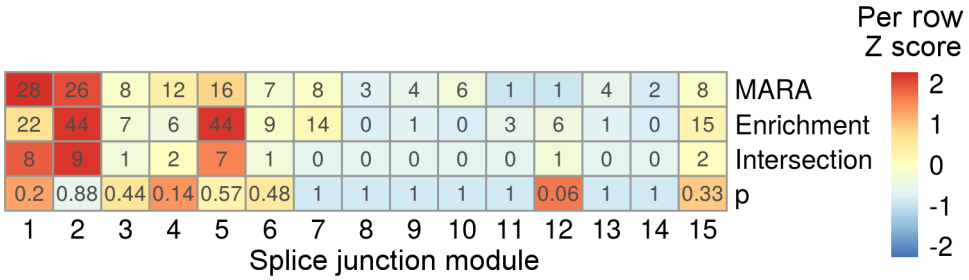

Figure 4-8. Numbers of splicing factor motifs associated with splice junction modules. MARA = numbers of splicing factor motifs with an activity correlated with each splice junction module eigenJunction PSI. Enrichment = numbers of splicing factor motifs with enriched counts flanking module splice junction sequences relative to "background" junctions. Intersection = numbers of splicing factor motifs identified in common by both approaches. p = probability of an intersection of the observed size or greater through random sampling (hypergeometric test).

4.3.4 Identifying candidate regulatory splicing factor motifs

Application of WGCNA revealed widespread and complex patterns of differential splicing across the CD4+ T cell activation timecourse, and involving genes from distinct biological processes. The regulatory splicing factors responsible for coordinating these patterns of differential splicing are of interest. To this end, S-MARA and splicing factor motif enrichment will be applied to generate a set of candidate regulatory splicing factors. This aim is similar to

that addressed in section 4.3.3, but is not specifically focused on associating splicing factor motifs to splicing modules. Rather, the aim is to identify motifs having the strongest associations with splicing variation across the timecourse using more focused approaches.

4.3.4.1 MARA-based predictions

To identify a higher confidence group of splicing factor motifs of interest, linear mixed effect spline modelling was performed at the individual motif level. Of the 103 motifs, 78 had significant relationships with time-after-activation (FDR < 0.05). To further refine this motif set, a more stringent filter of FDR < 5 x 10^-4 was used. The FDR was inversely correlated with the sample standard deviation of motif activity, and positively correlated with the range of motif activity values across the timecourse (a metric which acts as a motif activity effect size) (Figure 4-9). Filtering with this more stringent value therefore allowed identification of 62 motifs with large variation in activity across the timecourse and high consistency between donors (Figure 4-9). These 62 motifs were associated with 57 splicing factors collectively.

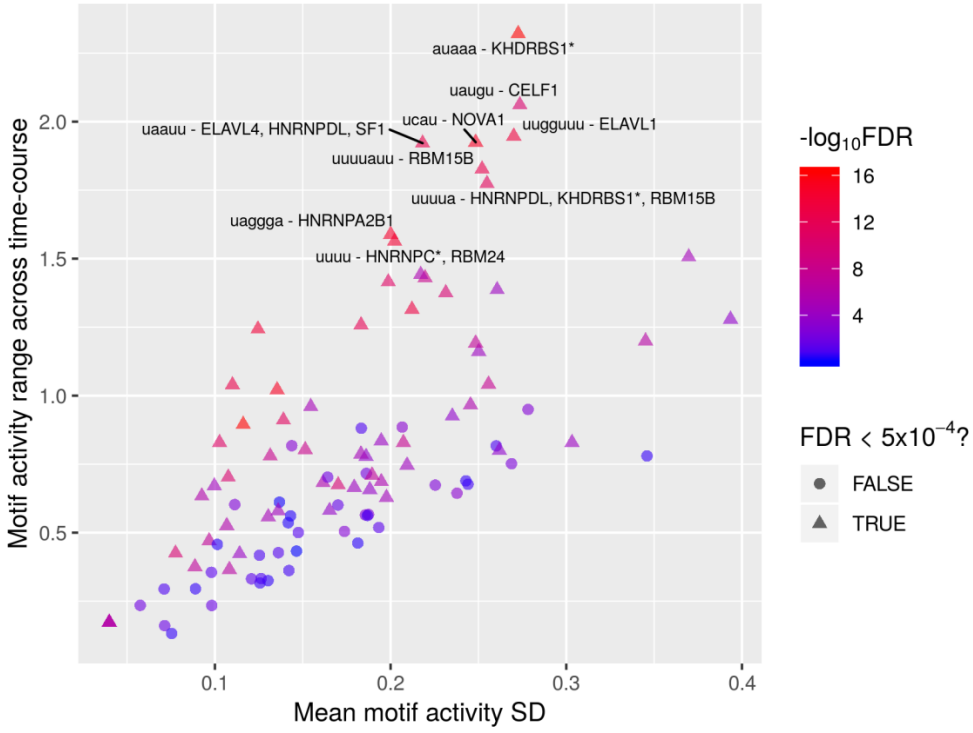

Figure 4-9. Characteristics of splicing factor motif activities across a timecourse of CD4+ T cell activation and polarisation. FDR = FDR of the probability of the null hypothesis (motif activity has no relationship with time after activation), assessed via linear mixed effect spline modelling. Motif activity range is the difference between maximum and minimum motif

activity across the timecourse. Mean activity standard deviation (SD) is the mean of SDs calculated across donors per time point and cell type. Motifs with an activity range > 1.5 are labelled with consensus motifs and associated splicing factors. * indicates a positive control splicing factor.

A number of motifs are known to promote binding of several different splicing factors, any of which may be driving the associated motif activity signal of interest. In such cases, gene expression data was used to further refine potential regulatory relationships between splicing factors and motifs. Nine of these 57 splicing factors of interest were either unexpressed or expressed at very low levels, and thus likely not driving the motif activity signal of their associated motif. Of the 62 selected motifs, 17 had an activity that correlated highly with the expression of one of the associated splicing factors (defined as absolute Pearson correlation > 0.55). In these instances, splicing factors without a correlative expression profile were no longer considered as potential regulators of these motifs. In cases where none of the associated splicing factors showed a correlative expression profile with a motif's activity, such filtering was not performed.

An illustrative example of the utility in this type of filtering is the GGGUA motif associated with both *PCBP2* and *HNRNPA2B1*. The activity of this motif correlated highly with *HNRNPA2B1* gene expression but not *PCBP2* expression (Figure 4-10), suggesting *HNRNPA2B1* is more likely to be the splicing factor driving the associated motif signal. Interestingly, amongst the motifs with increasing activity over time, this motif has the second strongest time-after-activation profile (as assessed via FDR), suggesting *HNRNPA2B1* may be a good candidate for a splicing regulator during the CD4+ T cell activation process. *HNRNPA2B1* has not previously associated with CD4+ T cell biology.

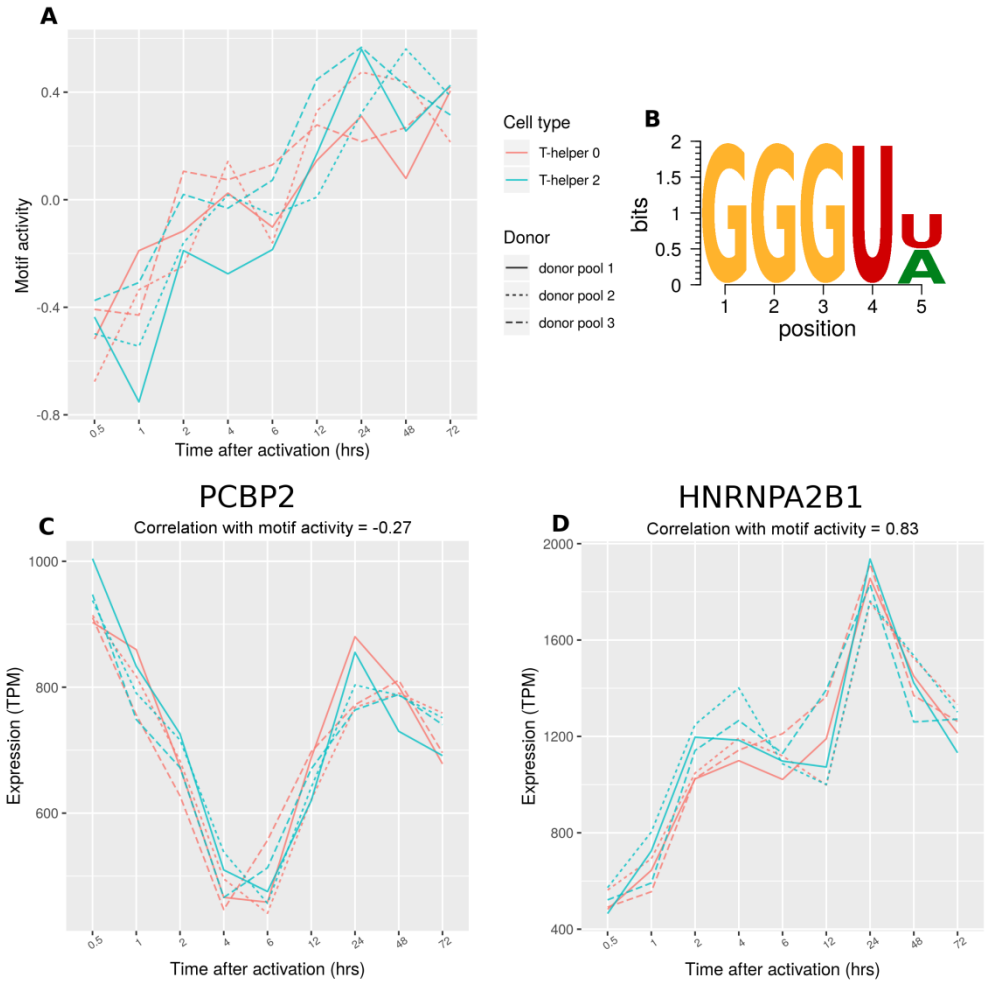

Figure 4-10. Motif activity and gene expression of *PCBP2* and *HNRNPA2B1* during CD4+ T cell activation and polarisation. (A) Activity of a *PCBP2-HNRNPA2B1* motif during CD4+ T cell activation and polarisation. **(B)** *PCBP2-HNRNPA2B1* motif logo. **(C)** *PCBP2* expression during CD4+ T cell activation and polarisation. **(D)** *HNRNPA2B1* expression during CD4+ T cell activation and polarisation. Donor pool = set of 12 technical replicates from an individual donor pooled for analysis after RNA-seq.

These filtering steps resulted in a final set of 47 candidate splicing regulators (depicted in Figure 4-5). To estimate the utility of these inferences, a set of positive control splicing factors was defined as those with previous evidence for a role in regulating differential splicing during T cell activation (Table 4-1). Of these 13 positive control splicing factors, 12 were in the final set of 47 candidate regulators. This represents a significant intersection as assessed via Hypergeometric test (probability of recovering 12 or more positive controls = 0.015). Prior to filtering based on gene expression there are 57 candidate splicing factors. Including the 12

positive controls, this is a less significant overlap (p = 0.18), and suggests value in incorporating gene expression information. Of note, the hnRNP LL CACACA motif did not have a significant relationship with time after activation (Figure 4-6A), despite hnRNP LL being a recognised splicing regulator during the CD4+ T cell activation process. Interestingly, the expression of *HNRNPLL* was relatively low (Figure 4-6A). The AUAAA motif associated with *KHDRBS1* had the lowest FDR for an association with time after activation. *KHDRBS1*, which encodes the Sam68 protein, shows an undulating pattern of expression with a net upregulation over time-after-activation; whilst the activity of the AUAAA motif decreases over time (Figure 4-6B). This negative correlation between gene expression and splicing activity may suggest a splicing repressor action of *KHDRBS1* through this motif. Of the motifs with a decreasing activity profile over time after activation, the *SFPQ* motif GUAGUGU had the lowest FDR. *SFPQ* expression initially increased before decreasing from the 2 hr time point (Figure 4-6C). This could be evidence of a negative feedback loop, where initial increases in expression increase the splicing activity of this factor, before expression levels are then decreased.

Table 4-1. Splicing factors with recognised regulatory roles during T cell activation, and with binding motif data available to facilitate motif analysis.

Splicing factor	Role in T cell activation	Alternative splicing demonstrated to act directly through binding *in cis* motifs?
HNRNPA1	Splicing of CD6 upon activation (demonstrated in bulk T cells) (Glória et al., 2014).	Yes
HNRNPC	Regulation of Mkk7 splicing during activation (demonstrated in Jurkat Cells) (Martinez et al., 2015).	Yes
HNRNPL	Splicing of CD45 in T cells in an activation responsive manner (demonstrated in JSL1 cells) (Rothrock et al., 2005). Regulates broader control of splicing in genes important to T cell development (work done	Yes

	in primary CD4+ T cells and JSL1 cells) (Cole et al., 2015; Shankarling et al., 2014).	
HNRNPLL	Global splicing regulator (including of CD45) during CD4+ T cell activation (Oberdoerffer et al., 2008).	Yes
HNRNPU	Splicing of MALT1 during CD4+ T cell activation (Meininger et al., 2016).	Yes
***KHDRBS1/*Sam68**	Splicing of CD44 during T cell activation (demonstrated in murine cell line – EL4) (Matter et al., 2002).	Yes
PTBP1	Regulation of gene expression upon T cell activation through multiple mechanisms including RNA degradation of IL-2 and CD40 transcripts (La Porta et al., 2016). Limited evidence for regulating CD45 splicing upon activation (Rothrock et al., 2005).	Limited evidence of a role in regulating alternative splicing during CD4+ T cell activation.
SFPQ/PSF	Splicing of CD45 during activation (demonstrated in Jurkat based model, follow-up work in primary CD4+ T cells) (Heyd and Lynch, 2010; Melton et al., 2007).	Yes
SRSF1	Splicing of CD45 exon 5 in an activation-responsive manner (Motta-Mena et al., 2010; Tong et al., 2005). Splicing of CD3 upon activation (Moulton and Tsokos, 2010). Splicing of CD6 upon activation (demonstrated in bulk T cells) (Glória et al., 2014; Lemaire et al., 1999).	Yes
SRSF2	Splicing of CD45 upon activation (demonstrated in murine T cells) (Wang et al., 2001).	Direct binding to CD45 RNA not demonstrated.
SRSF3	Splicing of CD6 upon activation (demonstrated in bulk primary T cells) (Glória et al., 2014).	Yes

TIA1	Regulation of Fas splicing in a Jurkat cell model (Izquierdo and Valcárcel, 2007). Recapitulates the splicing of Fas exon 6 inclusion seen upon activation of PBMCs (Liu et al., 1995).	Yes
U2AF2	Role as core spliceosome component, driving assembly of an RNA-protein interactome during CD4+ T cell activation (Whisenant et al., 2015).	Role in assembly of interactome and through binding constitutive splicing elements (the polypyrimidine tract).

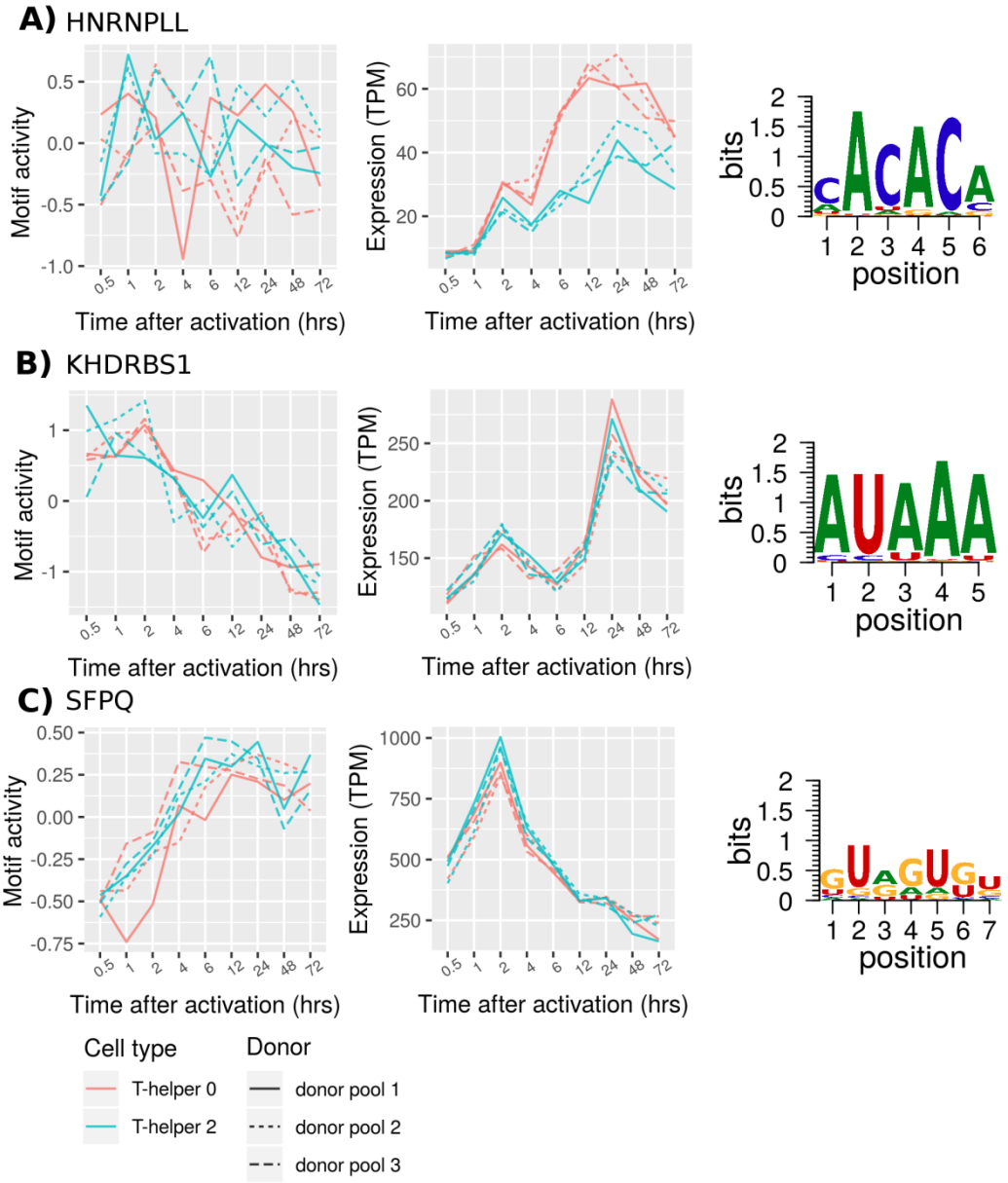

Figure 4-11. Motif activity, gene expression, and motif logos of selected splicing factors during CD4+ T cell activation and polarisation. (A) *HNRNPLL*, **(B)** *KHDRBS1* (Sam68), and **(C)** *SFPQ/PSF*. Donor pool = set of 12 technical replicates from an individual donor pooled for analysis after RNA-seq.

4.3.4.2 Motif enrichment-based predictions

In order to identify splicing factor motifs specifically associated with the splice junctions showing the strongest changes in splicing after CD4+ T cell activation, a motif enrichment procedure was tested. An initial differential splicing analysis comparing naïve cells with T_{h0}

cells at 12, 24, 48, or 72 hrs post-activation was performed. Differential splicing between naïve and activated cells was similar across all of these tested time points (Figure 4-12).

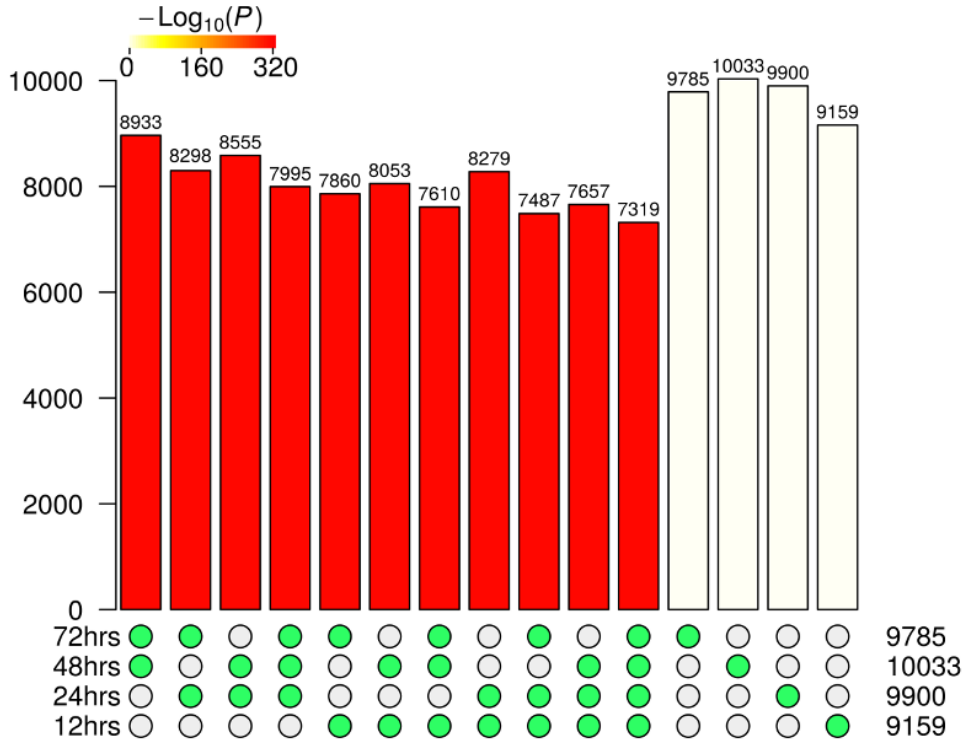

Figure 4-12. Intersections between differentially utilised splice junctions at various times after activation in T_{h0} cells. Green circles indicate the time points for which the intersection number in the above bar chart refer to. All intersections are significant as determined through Fisher's exact test.

The intersection of differentially utilised splice junctions across these four pair-wise time point comparisons (7319 splice junctions) was used for a motif enrichment analysis. This allowed identification of 35 splicing factor motifs with greater counts in RNA sequences flanking differentially spliced junctions relative to "background", non-differentially spliced junctions (one-tailed Wilcoxon rank sum test, FDR < 0.05). These 35 motifs are associated with 53 splicing factors, of which 10 were pre-defined positive controls (Table 4-1). The probability of an intersection of this size or greater is relatively high (p = 0.462, hypergeometric test). Of these splicing factors, 10 had low levels of gene expression and an additional one had expression that did not vary with time after activation (as assessed via linear mixed effect spline modelling, FDR > 0.05). Filtering of these 11 splicing factors led to an increased but non-

significant enrichment of positive controls within the candidate regulatory set (10 positive controls out of 42 post-filtering candidate splicing factors, p = 0.094, hypergeometric test). Of these 42 splicing factors, 35 are in the final set of 47 splicing factors identified as candidate activation regulators derived through the S-MARA based analysis, which is a highly significant intersection (probability of an intersection of this size or greater = 5.82 x $10^{\wedge-5}$ - hypergeometric test). Finally, the ten motifs with the greatest increase in motif counts relative to background junctions are depicted in Figure 4-13. These included the top MARA-based hit, *KHDRBS1* motif AUAAA, and additional positive control splicing factors *TIA1* and *U2AF2* (Figure 4-13).

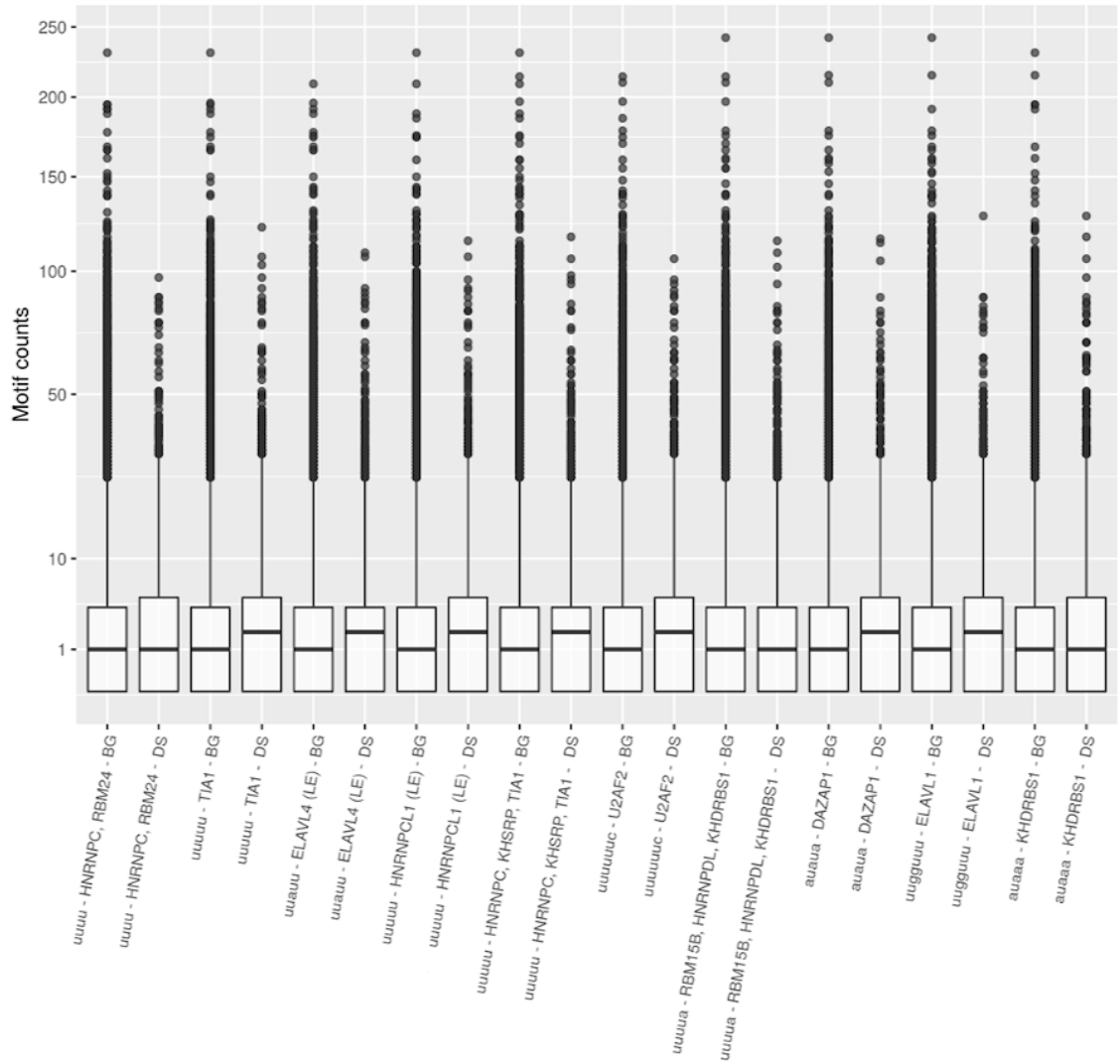

Figure 4-13. Splicing factor motif counts which are over-represented in splice junctions that are differentially spliced after CD4+ T cell activation. Motif enrichment was calculated using

the intersection of differentially spliced junctions at 12, 24, 48, and 72 hrs post-activation. The ten motifs with the highest increase in mean motif counts flanking differentially spliced (DS) junction sequences, relative to background (BG) junctions, are shown. Motifs are ordered from left to right in order of highest to lowest difference in mean motif counts. Motif consensus sequences and associated splicing factors are shown. To simplify labelling, and since multiple splicing factor can be associated with a given motif, factors with low expression levels are not shown, unless this resulted in removal of all splicing factors associated with a given motif (these cases are marked with LE – Low expression).

4.3.5 Gene-level splicing factor motif analyses may be biased towards identifying positive control factors

Thus far, the performance in identifying positive control splicing factors has been assessed via testing at the motif level, with S-MARA or motif enrichment analysis, before mapping to the gene level and employing hypergeometric enrichment testing. A limitation to this approach is that it does not account for the initial number of hypothesis tests performed relating to each splicing factor at the motif level. With this in mind, the numbers of motifs associated with each splicing factor in our compiled set 103 motifs was assessed. This revealed that positive control factors had significantly more motifs associated with them on average than non-positive control splicing factors. Specifically, positive control splicing factors had a median of three motifs, whilst non-positive control factors had a median of two motifs (Figure 4-14). This difference was significant as assessed via Wilcoxon rank sum test ($p = 0.018$). This feature of the data was not an intended characteristic of the compiled motif dataset. Indeed, since the results of the motif-based analysis conducted thus far are mapped to the splicing factor level in a post-hoc manner, the biased distribution in numbers of per-splicing factor motifs introduces a bias towards identifying positive control splicing factors.

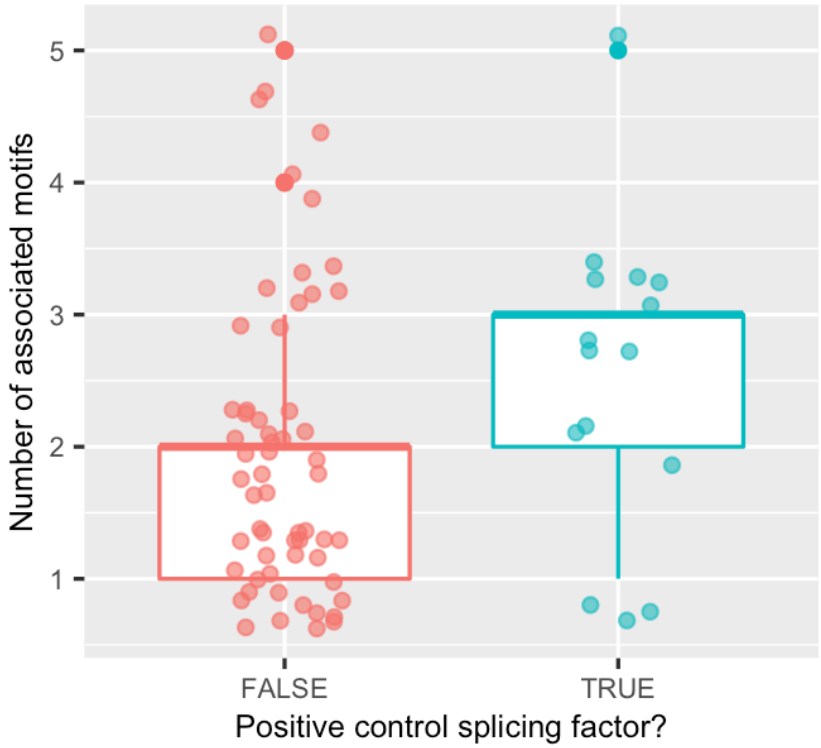

Figure 4-14. Distribution of number of associated motifs amongst positive control and non-positive control splicing factors. Splicing factors grouped on x-axis according to whether they are considered as positive control regulators during the CD4+ T cell activation process (Table 4-1).

4.3.6 Receiver operating characteristics of S-MARA and motif enrichment analysis in identifying positive control splicing regulators of CD4+ T cell activation

In light of the identified limitations in assessing performance of splicing factor motif analyses at the gene-level, an additional analysis was performed directly using motif-level data. To further assess the performance of the S-MARA and motif enrichment-based approaches in identifying positive control splicing factors, an assessment of the ROC AUC was performed. To this end, the per-motif FDR values derived from motif enrichment analysis (4.3.4.2) or S-MARA analysis (section 4.3.4.1) were used. Motifs were considered as true positives if associated with a positive control splicing factor, whilst all others were considered true negatives. This analysis identified the motif enrichment procedure as having greater performance characteristics. Indeed, although MARA showed an AUC > 0.5 (AUC = 0.561), the 95%

confidence intervals contained 0.5, whilst those of motif enrichment analysis did not (Figure 4-14).

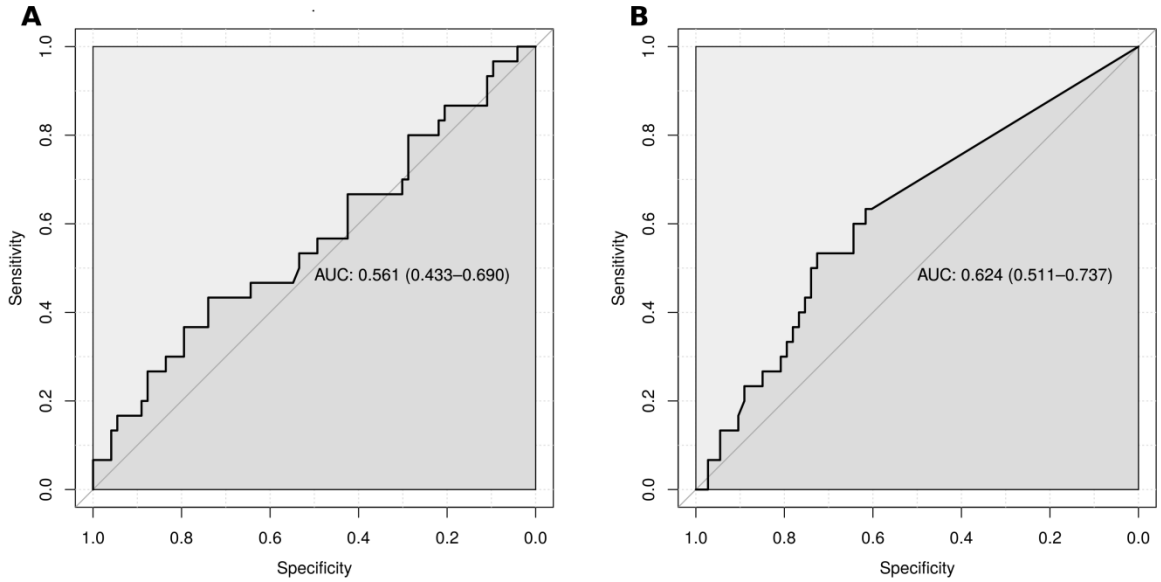

Figure 4-15. Receiver operating characteristics for the identification of positive control splicing factor motifs. (A) Performance of S-MARA. **(B)** performance of motif enrichment analysis. AUC = area under the curve. 95% confidence intervals are shown. Ratios of sensitivity to specificity in identifying positive control splicing factor motifs at varying FDR threshold values are depicted. True positives were defined as motifs associated with pre-defined positive control splicing factors (Table 4-1).

4.4 Discussion

Here, the regulation of splicing during the process of CD4+ T cell activation and polarisation to a T_{h2} subtype was studied through analysis of splicing factor motifs. To gain insight into the broad dynamics of splicing regulation, a correlation-network approach was used. Of the 103 splicing factor motifs studied, 78 showed an activity profile that varied over time after initial TCR stimulation (Figure 4-9) – behaviour consistent with potential roles in regulating activation-dependent alternative splicing. Clustering of these activity profiles resolved the motifs into groups of modules which potentially reflected different patterns of splicing regulatory control (Figure 4-4). Correlation-network analysis at the splice junction-level revealed that the predominant pattern of splicing variation was a simple and steady switch in relative splice junction usage (Figure 4-2, modules 1 & 2). Genes characterised by this splicing profile were enriched for roles in a wide range of biological processes representing most steps

of the gene expression pathway. These biological processes included splicing itself, in addition to other processes such as epigenetic modification and ribosomal biogenesis. This finding could reflect a role for alternative splicing in reprogramming gene expression to drive proliferation and production of key cytokines and associated receptors subsequent to CD4+ T cell activation. Indeed, CD4+ T cell activation is characterised by increased ribosomal biogenesis which is thought to facilitate increased cytokine production (Asmal et al., 2003). Further, the control of gene expression during the activation process is known to be influenced by several layers of epigenetics modifications (Schmidl et al., 2018). Across these modules of activity, the main source of variation in splicing was time after CD3/CD28 stimulation, rather than polarisation into a T_{h2} vs T_{h0} specification. However, a more directed analysis of differential splicing may highlight splicing modulations important to T_{h2} specification.

In order to infer which splicing factors may be driving the patterns of differential splicing observed across the identified splice junction modules, both a MARA-focused approach and a motif enrichment approach were applied. In this regard, the two methods produced largely non-overlapping predictions with regards to inferring potential splicing factor regulators of each splicing module (Figure 4-8).

To assess the relative performance of both S-MARA and motif enrichment analysis in identifying positive control regulatory splicing factors (Table 4-1), an additional analysis was performed. This additional approach focused on identifying patterns of strong splicing modulation across the timecourse, but unlike the previous analysis, did not rely upon identifying modules of distinct splice events. The S-MARA approach involved filtering motifs based on strength of association with time-after-activation, combined with a filtering of splicing factors based on gene expression profiles. This strategy resulted in a set of candidate regulatory splicing factors enriched for positive controls with known roles in regulation of alternative splicing during CD4+ T cell activation. Indeed, two of the top motif activity profiles of interest were for motifs associated with positive controls *SFPQ* and *KHDRBS1/*Sam68. Both of these splicing factors have been identified as regulators of splicing upon CD4+ T cell activation in the context of individual loci (Table 4-1). *KHDRBS1*/Sam68 showed a pattern of motif activity (motif = AUAAA) and gene expression consistent with a role as a splicing repressor (Figure 4-11). This contrasts with the documented role for Sam68 in promoting inclusion of CD44 exon v5 upon binding an AAAUU exonic sequence upon T cell activation

(Matter et al., 2002). However, Sam68 has also been shown to act as a splicing repressor in the context of spinal muscular atrophy (Pedrotti et al., 2010), and to have both repressor and enhancer activity during neurogenesis (Chawla et al., 2009). The strong association between activities of these binding motifs and splicing during the activation process may suggest a broad role in regulating the splicing of a larger number of genes.

Ranking splicing factor motifs by their relationship with time-after-activation in this manner can also highlight novel candidate regulators – genes without previously recognised roles in CD4+ T cell biology. For example, activity of the *HNRNPA2B1* motif GGGUA is strongly associated with time after CD4+ T cell activation, and highly correlated with *HNRNPA2B1* gene expression (Figure 4-10). A prediction from these results is that *HNRNPA2B1*, *SFPQ*, and *KHDRBS1*/Sam68 may have roles in regulating genome-wide programmes of alternative splicing during CD4+ T cell activation. This hypothesis could be tested via an experimental system in which splicing factor expression is reduced or overexpressed in CD4+ T cells exposed to an activation stimulus. These samples could then be re-analysed via RNA-seq to assess potential disruption to the activation-associated splicing regulatory programme.

To contextualize the results of S-MARA, splicing factor motif enrichment analysis was again applied. This approach, aimed at identifying motifs associated with the strongest changes in splicing after activation, identified a set of candidate splicing factors that significantly overlapped with the highest confidence MARA-based candidates. Both the MARA and motif enrichment methods, when combined with a filtering of splicing factors based on gene expression data, led to identification of a candidate list containing a number of positive controls with known roles in regulating splicing during CD4+ T cell activation. However, a potential source of bias towards recovering positive control splicing factors was identified, in that positive controls had more associated motifs (Figure 4-14). Positive control factors were thus more likely to be identified *ab initio*. In light of this, a ROC analysis was performed. Being conducted at the motif, rather than gene-level, this ROC analysis was not sensitive to the same bias, and is thus a more objective method of assessing performance. This direct assessment of per-motif scores revealed that motif enrichment derived results had improved specificity and sensitivity compared with the results produced through combining S-MARA with linear mixed effect spline modelling (Figure 4-14).

To give full consideration of these results, several limitations with the ROC analysis and the use of positive control splicing factors defined in Table 4-1 should be addressed. Firstly, one goal of applying splicing factor motif-based analyses in this study was to identify novel regulators of splicing during the CD4+ T cell activation and polarisation process. Indeed, several promising novel candidate regulators were highlighted through this analysis. This has obvious implications for the definition of true negatives when employing a ROC AUC analysis, whereby the splicing factors defined as true negatives will likely include a number of mislabelled cases of novel "positive case" splicing factor regulators. Similarly, defining true positives presents challenges. The strength of evidence for the pre-defined regulators of alternative splicing during CD4+ T cell activation is variable (Table 4-1). From extensively studied and high confidence cases such as hnRNP L, to factors such as hnRNP C which have only been studied in T cell lines rather than primary CD4+ T cells, or splicing factors for which only indirect evidence for a role in controlling alterative splicing during T cell activation has been demonstrated, such as PTBP1. Thus, there is a range of confidence with which each of the positive controls is indeed a key regulator of splicing during CD4+ T cell activation, and false positives may be present.

S-MARA displayed poor performance when applied to analysis of shRNA-induced splicing factor-knockdown data (Chapter 3). Initial application to the timecourse of CD4+ T cell activation herein proved more promising, since S-MARA identified a variety of splicing factor motif activity profiles that were consistent across replicates and may represent modular patterns of splicing regulatory activity (Figure 4-4). Further, identification of a subset of splicing factor motifs with high consistency across replicates and large variation across the timecourse (Figure 4-9) highlighted a group of splicing factors enriched for positive controls and including promising novel candidates. Issues of statistical power linked to low replicate numbers were identified as a possible cause of the poor performance in identifying regulatory factors in analysis of the ENCODE project data in Chapter 3. Thus, the greater numbers of biological replicates and the number of timepoints available for estimation of motif activity here may have aided the performance of S-MARA. However, as with the previous analysis, the ROC AUC of S-MARA was still inferior to the motif enrichment procedure (Figure 4-14). This finding thus warrants caution over interpretation of the results of S-MARA and highlights the need for experimental validation of novel predictions. In light of this, the predictions resulting from

application of motif enrichment analysis (Figure 4-13) are more promising and should be the focus of any potential follow-up investigations.

4.5 Conclusions

Herein, MARA was applied to infer splicing factor motif activity profiles across a timecourse of CD4+ T cell activation. Activity profiles which showed similar behaviour across replicates and were consistent with potential splicing regulatory profiles were identified. These profiles were filtered to produce a set of candidate splicing regulators. Although the results from application of S-MARA showed promise, analysis of differential splicing coupled with motif enrichment analysis again showed improved sensitivity and specificity characteristics. Thus S-MARA needs further refinement upon its current implementation before firm conclusions regarding the results of its application can be drawn.

Chapter 5. Assessment of the Genome-Wide Targets of the RNA Binding Protein Sam68 upon CD4+ T cell Activation

5.1 Introduction

KH domain containing, RNA binding, signal transduction associated 1 (*KHDRBS1*) encodes the multifunctional Src-associated in mitosis 68 kDa (Sam68) protein. Sam68 is a member of the Signal Transduction and Activation of RNA (STAR) family of proteins, which have roles in linking signal transduction pathways to post-transcriptional control of gene expression (Vernet and Artzt, 1997). The hnRNP K Homology (KH) domain of Sam68 confers RNA binding affinity towards several AU based motifs (Taylor and Shalloway, 1994). Sam68 has roles in multiple aspects of gene expression including transcription (Fu et al., 2013), alternative splicing (Matter et al., 2002; Paronetto et al., 2007), 3' end processing (La Rosa et al., 2016), and translation (Paronetto et al., 2009). The activity of Sam68 is modulated in response to extracellular signalling such as through stimulation of the TNF-alpha receptor (Kunkel and Wang, 2011) or T-cell receptor (Fusaki et al., 1997). In response, Sam68 acts as a scaffolding protein, interacting with the SH3 domain of various SRC kinases, including those downstream of TCR-signalling such as Fyn or Lck (Fusaki et al., 1997; Paronetto et al., 2003). Sam68 was identified as a regulator of CD44 exon v5 splicing, mediated through binding an *in cis* AAAAUU sequence, in response to T cell stimulation in the mouse T lymphoma EL4 cell line (Matter et al., 2002). In this context, Sam68 was proposed to be activated via phosphorylation subsequent to Ras-ERK signalling after T-cell stimulation (Matter et al., 2002).

A more widespread role for Sam68-regulated splicing during cellular differentiation processes has also been observed, specifically during neurogenesis (Chawla et al., 2009), adipogenesis (Huot et al., 2012), and spermatogenesis (Paronetto et al., 2011). In addition to its RNA-binding capacity, Sam68 mediates the activities of other proteins through direct protein-protein interaction. For example, the upregulation of CD25 upon TCR engagement was shown to depend upon Sam68 binding to the NF-κB complex, which in turn facilitates interaction with the CD25 promoter (Fu et al., 2013).

A genome-wide assessment of the role of Sam68 in transcriptional or post-transcriptional regulation of gene expression in CD4+ T cells has not been performed. During neurogenesis,

Sam68 was shown to regulate splicing of a specific set of pre-mRNAs enriched for Sam68-associated binding motifs (Chawla et al., 2009), and we hypothesise that this may also be true during the CD4+ T cell activation process. To investigate this hypothesis, *KHDRBS1* (Sam68) expression was knocked down in primary CD4+ T cells via RNA interference (RNAi), and RNA-seq was used to profile the transcriptome of both knockdown and wild type cells before and after activation. An activate, transduce, rest, reactivate protocol was employed to facilitate Sam68 knockdown in these primary CD4+ T cells (see Materials & Methods for details).

5.2 Aims

Differential splicing is widespread during the CD4+ T cell activation process. However, the underlying regulatory splicing factors controlling these events are unknown for many genes. Sam68 is a multifunctional RBP the activity of which is regulated through TCR signalling. We propose that Sam68 may contribute to the widespread regulation of differential splicing during CD4+ T cell activation. Using a genetic knockdown approach, we aim to:

1. Confirm knockdown of Sam68 in primary CD4+ T cells.
2. Identify genes with altered splicing or expression profiles induced via Sam68 knock down.
3. Intersect genes from aim 2 with genes also regulated specifically upon CD4+ T cell activation.
4. Assess enrichment of Sam68 binding motifs in genes with disrupted splicing in knockdown cells.

5.3 Results

5.3.1 Sam68 knockdown in primary CD4+ T cells

To functionally investigate Sam68, several experimental conditions were used - two control conditions (untransduced and scramble shRNA treated) and two knockdown conditions (Sam68 shRNA 1 & 2). Analysis of RNA-seq from these four conditions confirmed that Sam68 was depleted at the mRNA level (Figure 5-1) with an ~80% and ~56% reduction in expression in activated and resting cells respectively. PCA using either gene expression or splicing quantifications identified the first principal component of variance as being driven by activation state and the second by cell donor source (biological replicate) (Figure 5-2). Sam68 knockdown was associated with the fourth PC of variance (Figure 5-2).

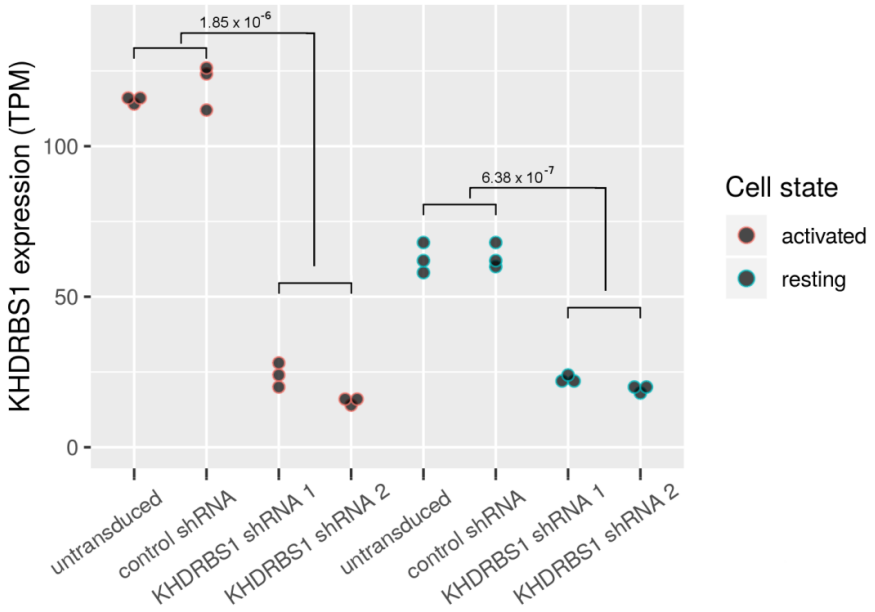

Figure 5-1. Sam68 mRNA expression in wild type and Sam68 knockdown CD4+ T cells. For hypothesis testing, control samples (untransduced and control shRNA), and knockdown samples (Sam68 shRNA 1 and 2) were pooled. FDR values for comparison of wild-type with Sam68 knockdown are depicted.

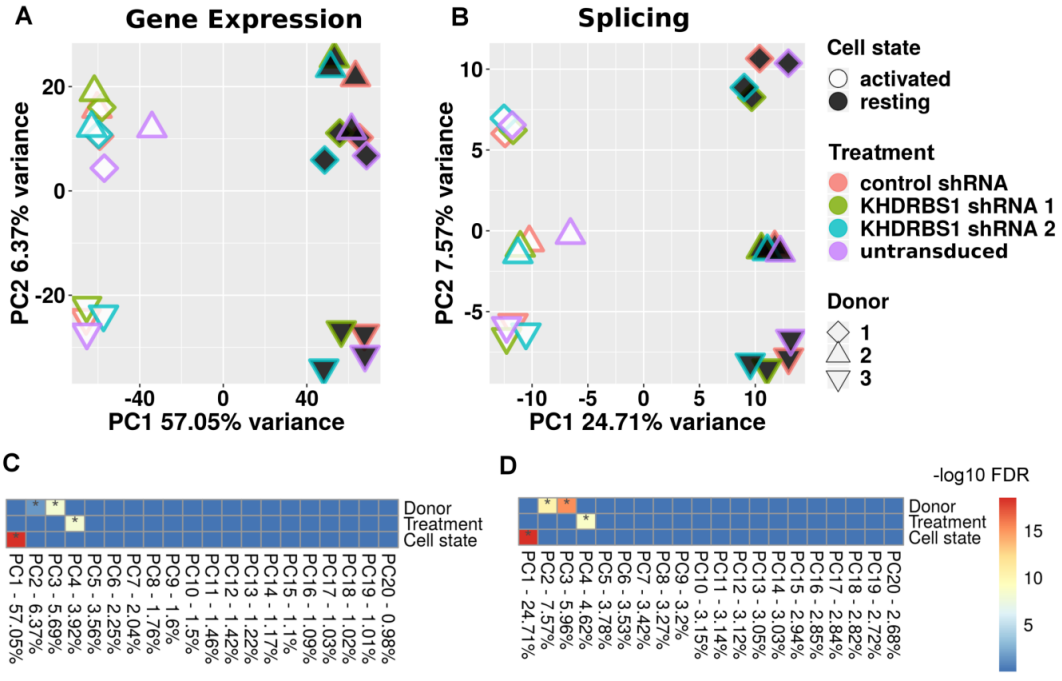

Figure 5-2. Principal component analysis of wild type and Sam68 knockdown CD4+ T cells. (A) & **(B)** Principal components of variance of : **(A)** Gene expression (regularized log transformed TPM), and **(B)** splicing (logit transformed PSI). **(C)** & **(D)** Linear regression of experimental

conditions against principal components of: **(C)** gene expression, and **(D)** splicing. * indicates a significant association.

In order to profile the regulation of gene expression upon CD4+ T cell activation and the effects of Sam68 knockdown, the three biological replicates from different biological conditions were compared in a pairwise manner (Table 5-1). Different control conditions (untransduced and scramble shRNA transduced) and knockdown conditions (Sam68-targeting shRNA 1 & 2) were pooled for this analysis. Differential mRNA abundance was assessed using Sleuth, and differential splicing via MAJIQ (See Methods for further details). As expected, differential gene expression and splicing analysis identified large numbers of regulated events upon CD4+ T cell activation (Figure 5-3A & D). Indeed, 72.9% of genes expressed in at least one sample (17633/24204) were identified as differentially expressed upon CD4+ T cell activation (FDR < 0.05). Sam68 knockdown also resulted in widespread differential gene expression in both activated (Figure 5-3B & E) or resting state cells (Figure 5-3C & F).

Table 5-1. Experimental conditions and pairwise comparisons used for differential gene expression and splicing analysis.

Reference condition	Comparison condition	Total sample number
Untransduced & control shRNA-treated – resting cells	Untransduced & control shRNA-treated - activated cells	12 (3 biological replicates per condition)
Control shRNA-treated & untransduced – resting cells	Sam68 targeting shRNA 1 & 2 – resting cells	12 (3 biological replicates per condition)
Control shRNA-treated & untransduced –activated cells	Sam68 targeting shRNA 1 & 2 – activated cells	12 (3 biological replicates per condition)

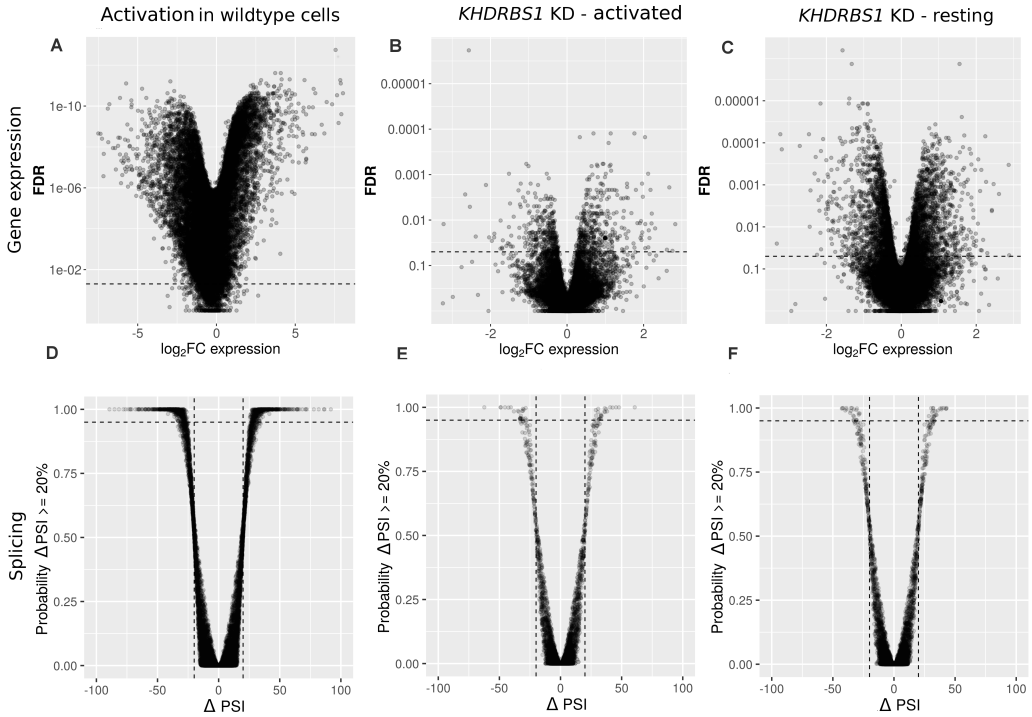

Figure 5-3. Volcano plots of differential splicing and gene expression in CD4+ T cells. (A-C) Differential gene expression. **(D-F)** Differential splicing. **(A & D)** activated control (untransduced & scramble shRNA transduced) samples relative to resting control samples, **(B & E)** Sam68 knockdown samples in resting state relative to control samples in resting state, **(C & F)** Sam68 knockdown samples in activated state relative to control samples in activated state. Comparisons as detailed in Table 5-1.

5.3.2 Sam68 dependent splicing during CD4+ T cell activation

In resting CD4+ T cells, 32 local splicing variants (LSVs) from 30 different genes were differentially spliced upon Sam68 knockdown (>= 95% probability of change in PSI >= 20% [see Appendix 8.4 for lists of these genes]). Upon activation, splicing of 30 LSVs from 28 genes was disrupted in Sam68 depleted cells relative to wild-type activated cell. The LSVs affected in the resting and activated states were largely independent, with only two LSVs disrupted in both cell states (Figure 5-4, Appendix 8.4). Activation of wild type CD4+ T cells stimulated altered splicing of 563 LSVs from 463 genes (Appendix 8.4), which overlapped modestly yet significantly with the set of Sam68 knockdown-sensitive LSVs in both resting or activated cell states (Figure 5-4). These gene sets were not enriched for any “Biological Process” gene ontology terms. Counter to previous findings in the mouse EL4 cell line (Matter et al., 2002),

splicing of *CD44* was not altered upon CD4+ T cell activation or knockdown of Sam68 (Figure 5-5).

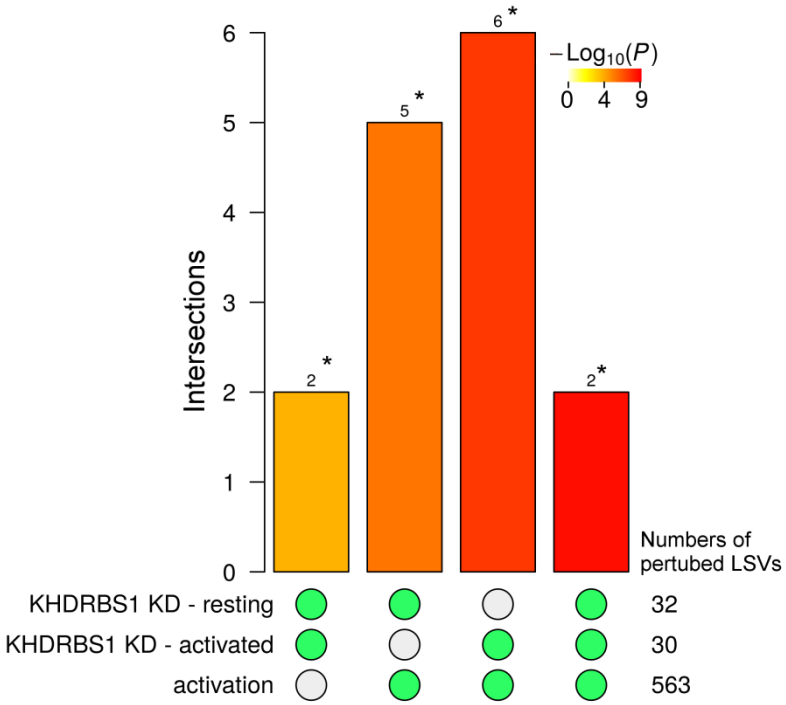

Figure 5-4. Intersections between local splicing variations with altered splicing upon Sam68 knockdown in CD4+ T cells. Green circles indicate the analyses for which the above intersection numbers relate to. * indicates cases with significant intersections relative to random sampling from the background set of all quantifiable LSVs, assessed via Fisher's exact test.

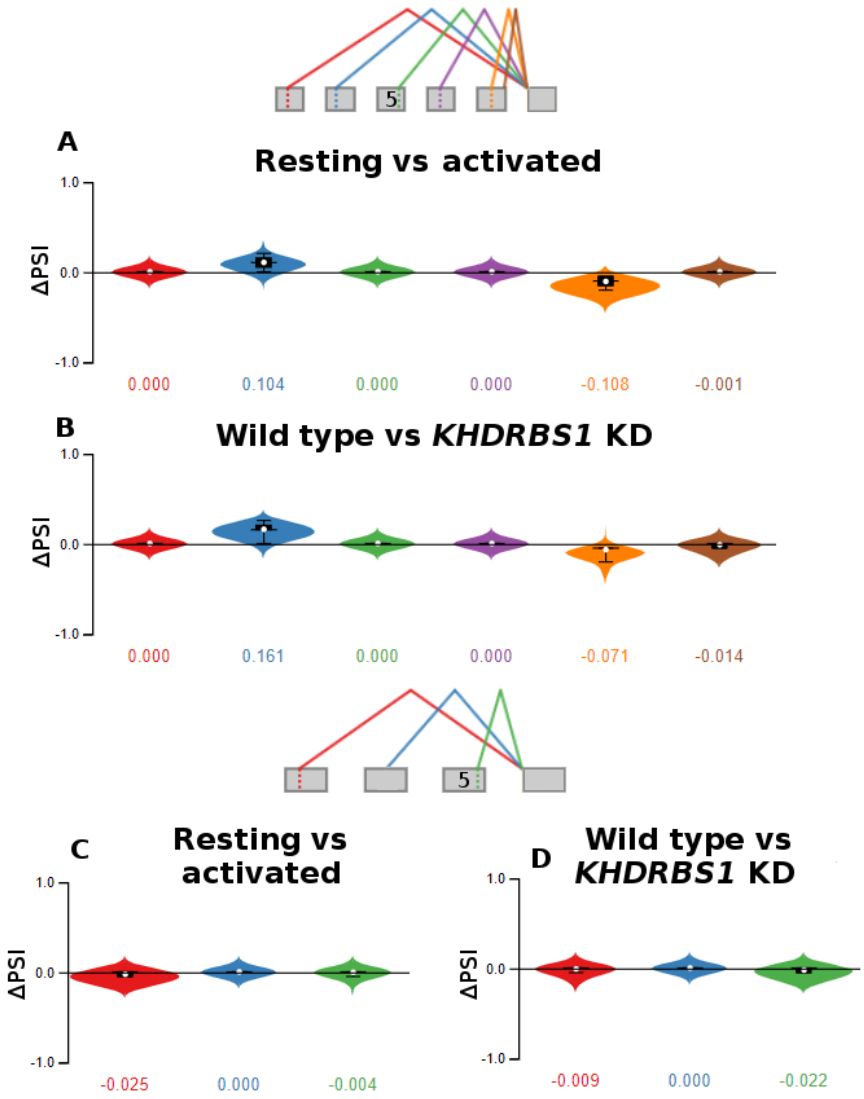

Figure 5-5. Splicing of CD44 exon v5 in two local splicing variations. Splicing of exon v5 is captured by two local splicing variations. **(A)** and **(C):** Splice junction usage in resting relative to activated CD4+ T cells. **(B)** and **(D):** Splice junction usage in wild type relative to Sam68 knockdown cells. Constitutive splicing of exon v5 with exon v4 not shown.

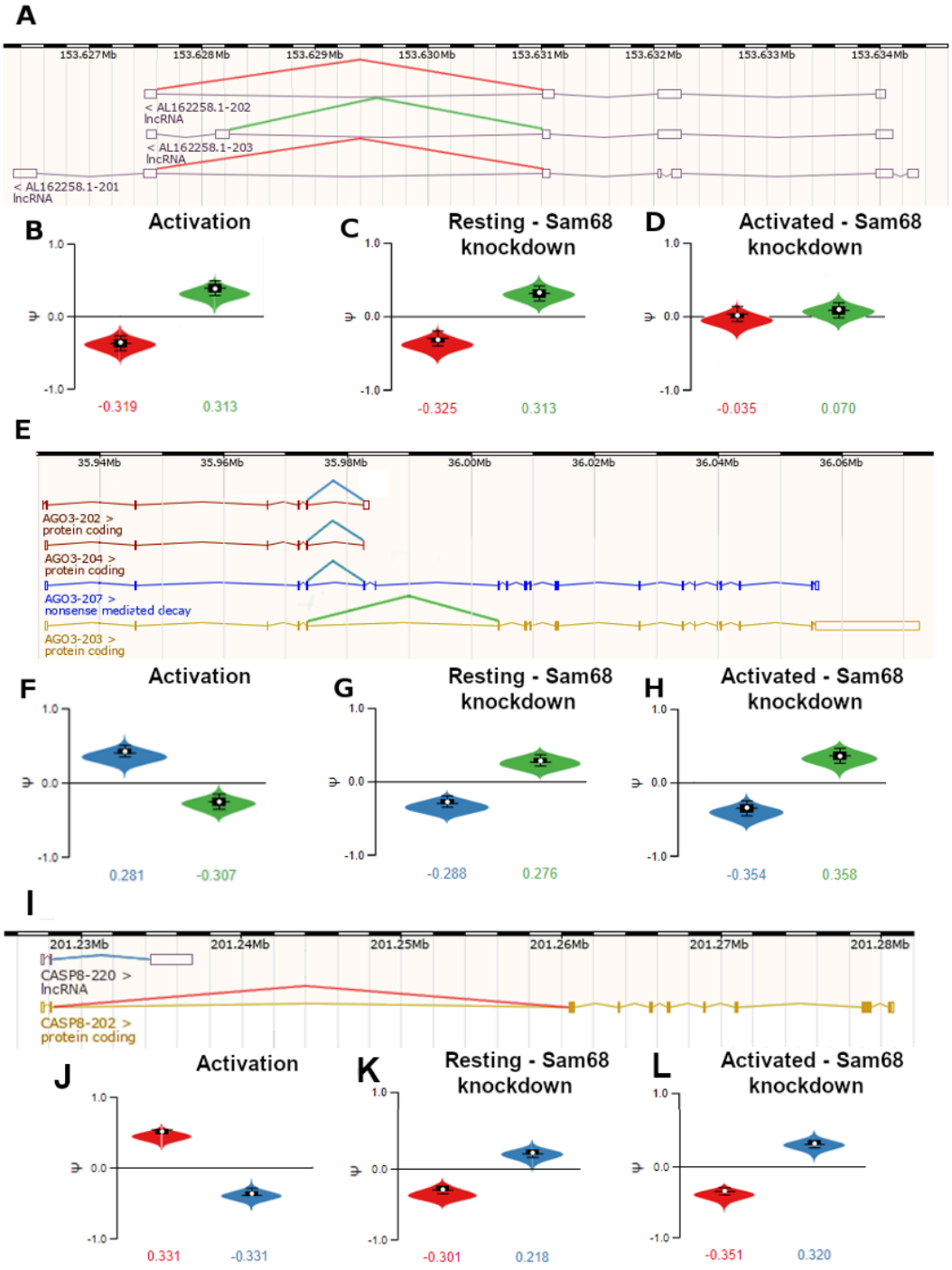

Figure 5-6. Two patterns of differential splicing upon CD4+ T cell activation and knockdown of Sam68. (A-D) Example of a locus at which Sam68 appears to promote a resting cell-like splicing profile. **(A)** Schematic shows gene structure and isoforms of *AL162258.1*, an antisense lncRNA. **(B-D)** change in PSI of two splice junctions: **(B)** upon activation, **(C)** upon knockdown of Sam68 in resting cells, or **(D)** knockdown of Sam68 in activated cells. Colours relate to splice junctions depicted in **(A)**, the red splice junction is used in two annotated transcripts. **(E-H)**

Example of a locus at which Sam68 appears to promote an activation-associated splicing profile. **(E)** Schematic of *AGO3* isoforms. **(F-H)** change in PSI of two splice junctions: **(F)** upon activation, **(G)** upon knockdown of Sam68 in resting cells, or **(H)** knockdown of Sam68 in activated cells. Colours relate to use of splice junctions depicted in **(E)**. **(I-L)** Example of a locus at which Sam68 appears to promote an activation-associated splicing profile. **(I)** Schematic of *CASP8* isoforms. **(J-L)** change in PSI of two splice junctions: **(J)** upon activation, **(K)** upon knockdown of Sam68 in resting cells, or **(L)** knockdown of Sam68 in activated cells. Colours relate to use of splice junctions depicted in **(I)**. Isoform schematics adapted from Ensembl genome browser release 97 (Hunt et al., 2018).

Addressing whether Sam68 has a functional contribution to differential splicing during CD4+ T cell activation is a priority. Six LSVs were differentially spliced both upon activation and after Sam68 knockdown, providing evidence for a potentially minor role of Sam68 in controlling activation-associated differential splicing (Table 5-1). Several genes displayed patterns of differential splicing upon activation that were also promoted upon Sam68 knockdown, namely *AL162258.1* (Figure 5-6A-D), *GOLGA8B*, and *TMEM116*. This suggests Sam68 acts to maintain the resting state-associated splicing profiles at these loci. Splicing at *AGO3* (Figure 5-6E-H), *CAPS8* (Figure 5-6I-L), and *RRP7BP* showed a different response in which Sam68 knockdown inhibited the activation-associated alternative splicing pattern. At these loci, Sam68 may thus act to enhance splicing modulation during the T-cell activation process. Examining the gene models of several of these differentially spliced loci suggests that Sam68 knockdown may promote a change in isoform usage (Table 5-1). Some instances of differential splicing involved the use of novel, unannotated exons or splice sites (*RRP7BP* and *GOLGA8B* – Table 5-1).

Table 5-2. Effects of CD4+ T cell activation and Sam68 knock down on splicing of selected exons and corresponding transcript isoforms. Ensembl transcript identifiers and associated biotypes are shown (Ensembl version 97). Co-ordinates of alternative and differentially spliced exons relate to GRCh38/hg38 assembly. lncRNA = long non-coding RNA.

Gene	Isoform/s upregulated in wild type cells upon activation	Isoform/s upregulated in Sam68 knockdown relative to wild type cells	Change in PSI of exon/splice junction upon Sam68 knockdown

AGO3	ENST00000324350.9, ENST00000397828.3 – protein coding	ENST00000373191.9 – protein coding	29%
CASP8	ENST00000490682.5 – processed transcript	ENST00000264274.13– protein coding	-30%
RRP7BP	NA – use of novel splice acceptor: chr22:42571883	ENST00000437211.5 - processed transcript	31%
AL162258.1	ENST00000472233.1 – lncRNA	ENST00000472233.1- lncRNA	33%
GOLGA8B	ENST00000342314.9 – protein coding	ENST00000342314.9 – protein coding	36%
TMEM116	ENST00000550831.7 – protein coding	ENST00000550831.7 – protein coding	31%

Motif enrichment analysis was performed to assess the distribution of Sam68-associated motif occurrences in regions flanking differentially spliced events. Splice junctions with altered splicing upon activation had greater counts for Sam68 motifs AUAAA and UUUUA in their flanking RNA sequences relative to background junctions (one tailed Wilcoxon rank sum test, FDR = 0.009). Junctions affected by Sam68 knockdown in resting cells, but not those affected in activated cells, also showed greater counts for the Sam68 AUAAA motif (one tailed Wilcoxon rank sum test, FDR = 0.043) (Figure 5-7).

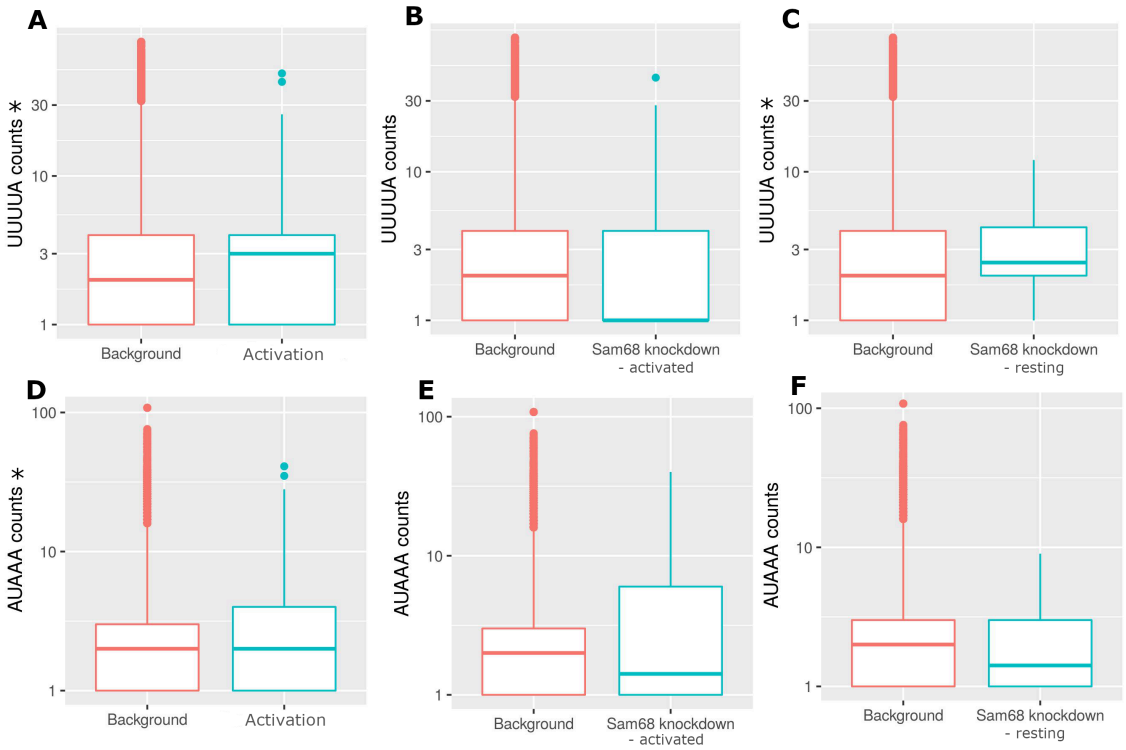

Figure 5-7. Motif enrichment analysis for Sam68-associated motifs using differentially regulated splice junctions. (A-C) Count distributions for motif UUUUA amongst differentially spliced junctions after: **(A)** CD4+ T cell activation, **(B)** depletion of Sam68 mRNA in activated cells, **(C)** depletion of Sam68 in resting cells. **(D-F)** Count distributions for motif UUUUA amongst differentially spliced junctions after: **(D)** CD4+ T cell activation, **(E)** depletion of Sam68 mRNA in activated cells, **(F)** depletion of Sam68 in resting cells. Background = count distribution amongst non-differentially regulated background junctions. * indicates the differentially regulated junctions were over-represented for motif counts relative to background splice junctions.

5.3.3 Sam68-dependent gene expression during CD4+ T cell activation

Sam68 can regulate multiple steps in the gene expression pathway. Therefore, a differential gene expression analysis was performed to assess the effects of Sam68-depletion on gene-level mRNA abundances. This analysis identified 3758 genes with altered expression upon Sam68 knockdown in resting CD4+ T cells (Figure 5-3B). Upon activation of Sam68 depleted cells, 675 genes had altered expression relative to wild-type activated cells (Figure 5-3C). A significant overlap between the genes affected in both the resting and activated state was present (Figure 5-8). A number of genes differentially expressed in Sam68 knockdown cells also displayed altered splicing (Figure 5-8). Counter to previous findings (Fu et al., 2013), *IL2RA*

(CD25) expression was not reduced by Sam68 knockdown (Figure 5-9A). Gene ontology analysis revealed that Sam68 regulated genes in resting cells were enriched for a role in "DNA strand elongation involving in DNA replication", whilst those genes affected in the activated cells state were enriched for roles in "catecholamine metabolic process", "T cell costimulation" (including *CD28* – Figure 5-9B), and "interferon-gamma-mediated signalling pathway".

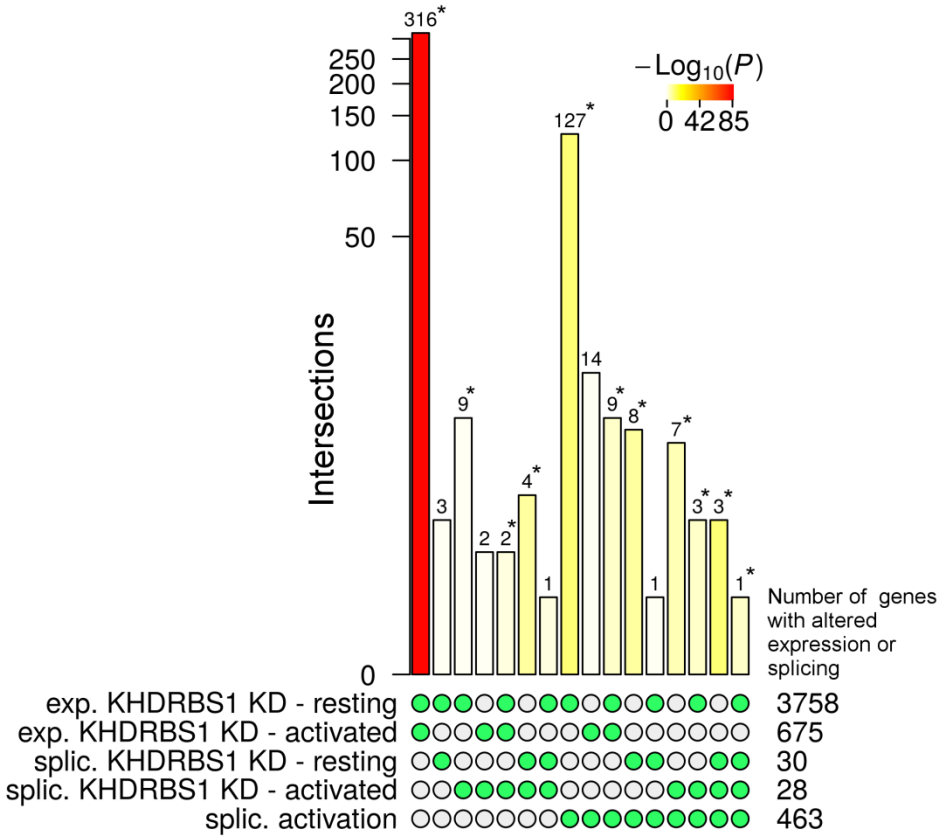

Figure 5-8. Intersections between genes with altered expression or splicing upon activation or Sam68 knockdown in CD4+ T cells. exp. = altered gene expression, splic. = altered gene splicing. KD = knockdown. Green circles indicate the analyses for which the above intersection numbers relate to. * indicates cases with significant intersections relative to random sampling from the background set of all expressed genes, as assessed vis Fisher's exact test.

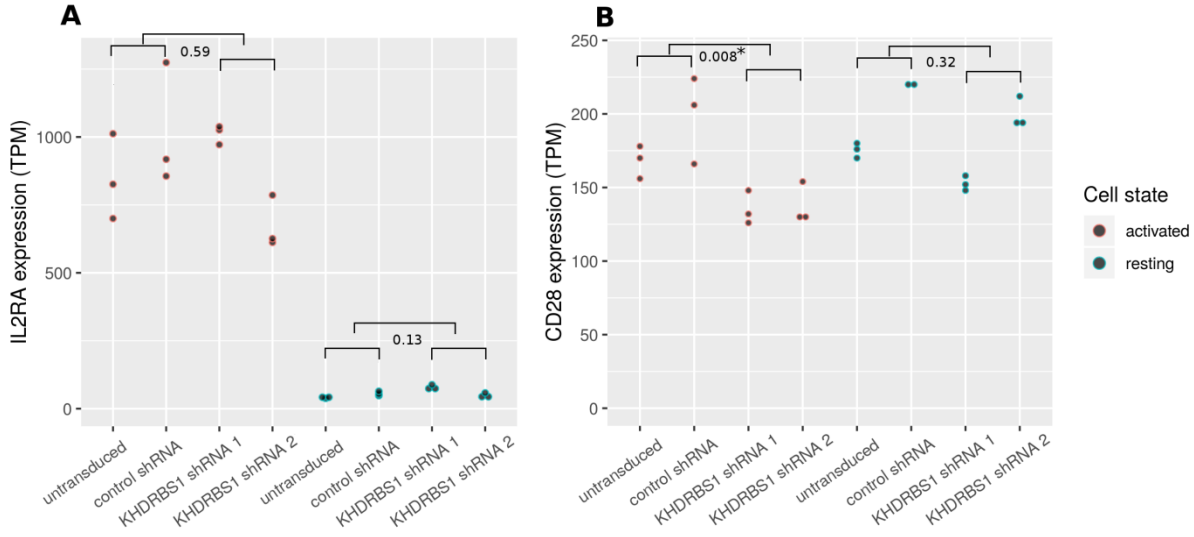

Figure 5-9. Gene expression of selected cell surface marker genes in wild type and Sam68 knockdown CD4+ T cells. (A) *IL2RA* (CD25) and **(B)** *CD28*. FDR shown, * highlights FDR < 0.05. For hypothesis testing, control samples (untransduced and control shRNA treated), and knockdown samples (Sam68 shRNA 1 and 2 treated) were pooled.

5.3.4 Comparison of activation and re-activation induced splicing in CD4+ T cells

In this study, CD4+ T cells were stimulated using a re-activation procedure wherein CD3/CD28 stimulation was initially applied to facilitate transduction with lentiviral vectors containing Sam68-targeting shRNA. The activation stimulus was then removed to allow cells to return to a resting state for three days, before being reactivated. This method contrasts with some previous investigations of CD4+ T cell activation. For instance, Henriksson *et al.* (Henriksson et al., 2019) activated naïve CD4+ T cells via CD3/CD28 stimulation followed by IL-2 exposure, whilst Ni *et al*. (Ni et al., 2016) studied CD4+ T cells at 18 hrs post-activation via CD3/CD28 stimulation. These studies therefore investigated differential splicing in true naïve CD4+ cells exposed to a "primary" activation stimulus, rather than after a secondary activation exposure (re-activation). Different activation protocols may produce differing effects on alternative splicing. We therefore compared the splicing modulation induced through these various activation protocols.

Large and highly significant intersections were observed between the sets of genes with altered splicing at 12, 24, 48, or 72 hrs post-activation in the study from Henriksson *et al.* and

those with altered splicing at 18 hrs post-activation in the Ni *et al.* study (Ni et al., 2016) (Figure 5-10). Indeed, the majority of differentially spliced genes identified in the Henriksson *et al.* data are also observed in the Ni *et al.* data. The Ni *et al.* genes are thus essentially a superset of the Henriksson *et al.* genes. In contrast, pairwise intersections between the genes with altered splicing upon re-activation (this study) and those regulated at 12 or 24 hrs post-activation (Henriksson study) were smaller but statistically significant (Figure 5-10). The genes alternatively spliced at 18 hrs post-activation (Ni study) also significantly overlapped with the re-activation sensitive gene set, with a somewhat larger overlap of 195 genes (Figure 5-10). Therefore, although there is similarity between differential splicing in CD4+ T cells in response to primary activation vs secondary re-activation, two independent experiments investigating splicing directly after a primary activation stimulus showed much greater similarity.

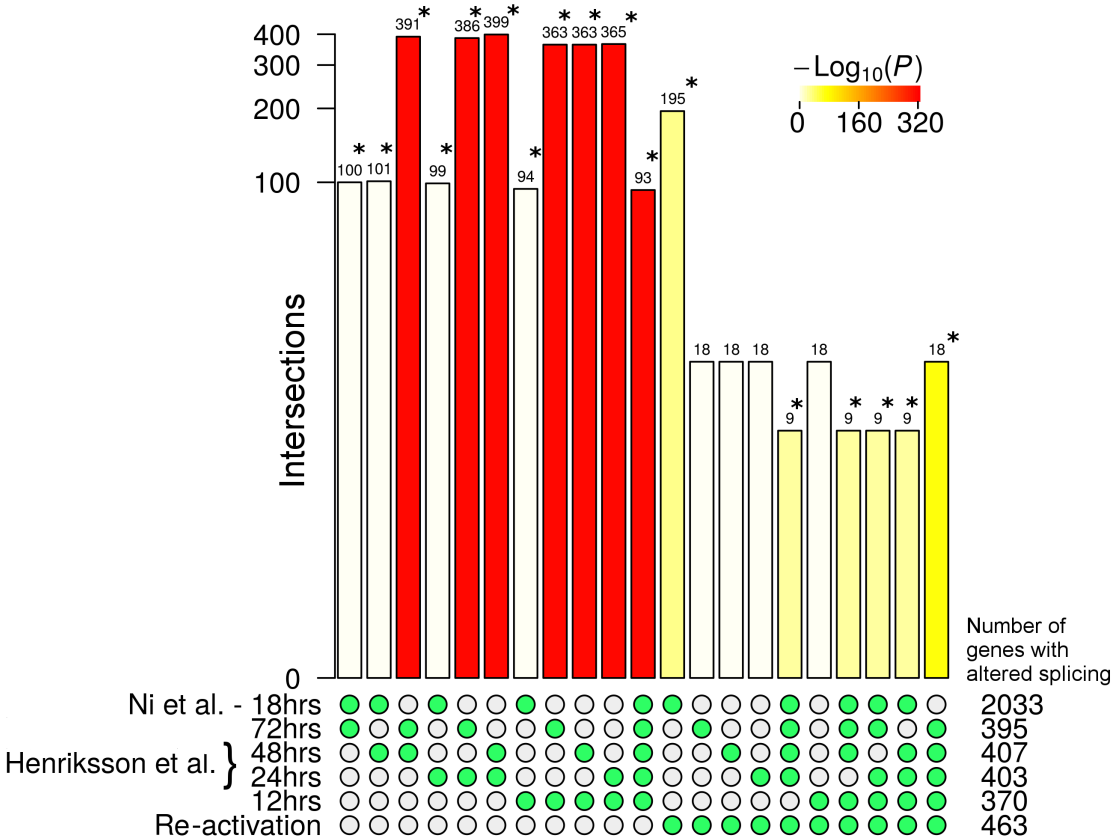

Figure 5-10. Intersections between genes with altered splicing upon CD4+ T cell activation or re-activation after various time points and from multiple studies. Genes with altered splicing at: 18 hrs post-activation (Ni *et al.* – 18 hrs), 12, 24, 48, or 72 hrs post-activation (Henriksson *et al.*), or 3 days after re-activation (this study, Re-activation). Green circles indicate the analyses for which the above intersection numbers relate to. * indicates cases with significant

intersections relative to random sampling from the background set of expressed genes, as assessed via Fisher's exact test.

In line with the data generated from this study (Figure 5-5), CD44 exon v5 was not differentially spliced upon activation in the Ni *et al.* or Henriksson *et al*. studies (Figure 5-11). This suggests that, although Sam68 has been previously reported to regulate splicing of CD44 in response to T-cell stimulation (Matter et al., 2002), this may be a consequence of the experimental system used. Specifically, Matter *et al.* (Matter et al., 2002) utilised the mouse T lymphoma EL4 cell line to study Sam68-mediated alternative splicing.

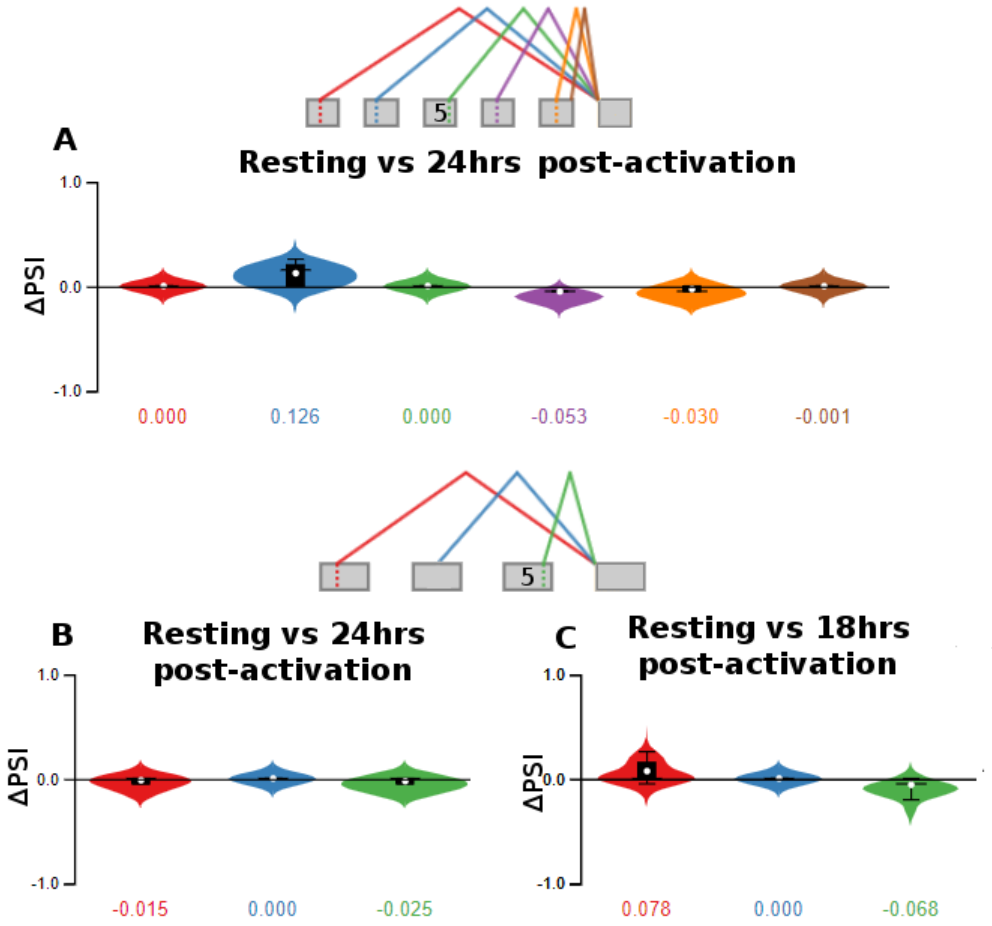

Figure 5-11. Alternative splicing of CD44 exon v5 in two local splicing variations. (A) & **(B)**: splice junction usage in CD4+ T cells 24 hrs post-activation relative to resting cells – data from Henriksson *et al.*. **(C)** Splice junction usage in CD4+ T cells 18 hrs post-activation relative to resting cells – data from Ni *et al.*. Splice junctions from the LSV in **(A)** were not utilised in RNA-seq data from the Ni *et al.* experiment, instead the splice junction shown in brown was

constitutively utilised. Exon v5 is spliced with exon v4 but this is a constitutive event not captured in an LSV.

5.4 Discussion

Sam68 is recognised as a regulator of multiple steps in the gene expression pathway during T-cell activation. Here, RNAi-mediated silencing was used to investigate the genome-wide splicing targets of Sam68. Gene expression was successfully reduced at the mRNA level in both the resting and activated CD4+ T cell state (Figure 5-1). Several thousand genes had altered RNA abundance after knockdown of Sam68, and a smaller number showed altered patterns of alternative splicing (Figures 5-3 & 5-8). The splice junctions with altered patterns of use in Sam68 knockdown cells in the resting state were over-represented for the Sam68 motif AUAAA in their flanking RNA sequences (Figure 5-7). This suggests a potentially direct mechanism of splicing regulation involving Sam68 binding to *in-cis* elements.

Several splicing events that were altered in Sam68 knockdown cells appeared to control switches in expression of distinct annotated isoforms (Table 5-1). This included *CASP8*, which was differentially spliced upon CD4+ T cell activation to favour expression of a protein coding transcript over an alternative lncRNA (Figure 5-6I-L). Knockdown of Sam68 in both the resting and activated state altered the splicing profile to favour expression of the resting state-associated lncRNA coding variant. This suggests a role of Sam68 in either repressing usage of the protein coding isoform or promoting usage of the non-coding transcript in an activation-associated manner. Potential functions of the non-coding transcript are unknown, but the protein coding product caspase-8 has been shown to be necessary for T-cell development and activation in the mouse (Salmena et al., 2003). Sam68 has been previously shown to promote the apoptotic activity of caspase-8 via recruitment of RIP to the FADD-caspase-8-complex which is necessary for caspase-8 activation (Ramakrishnan and Baltimore, 2011). However, this process is mediated through protein-protein interactions and is independent of the RNA-binding and splicing function of Sam68. Other Sam68-sensitive splicing events were related to the use of unannotated splice junctions, as with *GOLGA8B* (Table 5-1). This set of activation-sensitive, Sam68-regulated splicing events could be selected for future study. For instance, *in silico* analysis could be employed to predict whether the induced patterns of alternative splicing may alter protein coding sequences, such as via introduction of premature termination codons or functional protein domains. Future experimental work could include functional

investigation through over-expression or knockdown assays of specific Sam68-regulated isoforms. Further, although only six genes were differentially spliced both in response to activation and Sam68 knockdown, splicing of additional genes was altered specifically in response to Sam68 knockdown (Figure 5-4), and these events may also be of interest for future study.

Knockdown of Sam68 resulted in differential splicing of ~30 genes. Counter to expectations, *CD44* was not amongst these genes. The role of Sam68 as a signaling-dependent splicing modulator was initially established through work showing that exon v5 of *CD44* is differentially spliced in response to T cell stimulation (Matter et al., 2002). However, this exon was not differentially spliced upon Sam68 knockdown in our experiments (Figure 5-6). However, *CD44* exon v5 was also not found to be differentially spliced upon CD4+ T cell activation in data from several other studies examined herein (Henriksson et al., 2019; Ni et al., 2016) (Figure 5-5 & Figure 5-11). This discrepancy may be due to differences in the cell models utilised. The study linking Sam68 with *CD44* splicing was conducted in a mouse T-lymphoma line (EL4 cells). In one study, expression of *CD44* variable exons (including v5) was specifically observed in lymphocytes from patients with lymphoma, but not from healthy donors (Khaldoyanidi et al., 1996). These observations suggest that Sam68-dependent splicing of *CD44* may then be specific to the cancer-induced cell environment.

In our previous analysis, the activity of the Sam68 binding motif AUAAA was strongly associated with splicing modulation across a timecourse of CD4+ T cell activation, and was over-represented in events that were differentially spliced upon activation (Chapter 4, data from Henriksson *et al.* (Henriksson et al., 2019)). One interpretation of these finding is that Sam68 may play a widespread role in splicing regulation upon CD4+ T cell activation. However, knockdown of Sam68 resulted in a relatively small set of genes with perturbed splicing in CD4+ T cells after re-activation, and these genes were not enriched for roles in distinct biological processes. This suggests that in fact Sam68 has only a minor role in regulating activation-induced differential splicing. Several important caveats to this conclusion should be considered however. Firstly, in this study Sam68 expression was successfully reduced (Figure 5-1), but a complete knockout was not performed, and may have resulted in greater alterations to cellular alternative splicing regulation. Redundancy effects may also be a contributing factor,

whereby other splicing factors with partially overlapping functionality and binding preferences to Sam68 could compensate for the reduced Sam68 splicing activity in knockdown cells.

In addition, the experimental activation protocol employed should be considered. In this study, a rest, activate, rest, re-activate protocol was used to facilitate RNAi-mediated Sam68 depletion. Thus, the functions of Sam68 were investigated in previously activated cells exposed to a second activation stimulus. Two independent studies of naïve CD4+ T cell primary activation showed a high similarity in activation-induced differential splicing (Figure 5-10). In contrast, the re-activation-induced splicing pattern identified herein showed less similarity (Figure 5-10), suggesting that the secondary activation response may only partially reflect the transition from naïve to activated CD4+ T cells. Indeed, previous work has found that expression of some cell surface markers, including CD45 variants, was less responsive to stimulation of CD4+ T cells that had been previously exposed to antigen (i.e. re-activation) (Brenchley et al., 2002).

Sam68 knockdown resulted in a greater number of changes to gene expression (i.e. mRNA abundances) than to splicing. This included regulation of genes enriched for roles in T cell co-stimulation, such as a down-regulation of *CD28* (Figure 5-9). Modulation of *CD28* expression has been documented in response to TCR stimulation and *CD28* co-stimulation as part of a negative feedback response (Vallejo et al., 1999). Sam68 is known to contribute to transcriptional regulation through interacting with NF-κB (Fu et al., 2013), but has not been previously associated with *CD28* expression. Whilst the focus of this study was differential splicing, the widespread changes to RNA abundance in Sam68 knockdown cells are interesting and could be further investigated. For instance, we may hypothesise that the changes in gene expression upon Sam68 knockdown are mediated through modulation to NF-κB activity. One option for investigating this hypothesis further could be to use MARA to estimate NF-κB motif activity in wild-type compared with Sam68 knockdown cells. Down-regulation of NF-κB motif activity would then provide initial support to this hypothesis which could be investigated further. This application of MARA would be as per the published and validated use for the inference of transcription factor activity (Balwierz et al., 2014; Madsen et al., 2018).

5.5 Conclusions

Sam68 was investigated for its genome-wide contributions to gene expression and splicing during the CD4+ T cell activation process. A number of genes that were differentially regulated upon activation were also affected by Sam68 depletion, and these could be investigated further to determine the potential functional consequences of the specific Sam68-associated splicing modulations. The results of this study suggest that Sam68 regulates gene expression in CD4+ T cells through control of both RNA abundance and alternative splicing, but that the role in control of alternative splicing may be less widespread.

Chapter 6. Alternative Splicing of HIV-1 Transcripts is Disrupted by Introduction of CpG Dinucleotides to the Viral Genome

6.1 Introduction

6.1.1 CpG suppression in the HIV-1 genome facilitates evasion of host restriction

HIV-1 possesses a ~9kb genome which encodes 15 distinct proteins (Watts et al., 2009). After viral integration, the proviral DNA is processed through the host gene expression pathway to facilitate transcription, 5' capping, splicing, polyadenylation, nuclear export, and translation (Karn and Stoltzfus, 2012). In addition to encoding for viral proteins, the HIV-1 genomic RNA contains conserved *cis*-acting elements which facilitate these various steps in the viral lifecycle (Mayrose et al., 2013; Ngandu et al., 2008). Well characterised elements include the RRE located in *env* (Pollard and Malim, 1998), splicing signals within *pol, vif, vpr, tat, rev,* and *env* (Stoltzfus, 2009), the polypurine tracts within *nef* and *pol* which facilitate reverse transcription (Le Grice, 2012), and the ribosomal frameshift element within *gag* necessary for Pol translation (Brierley and Dos Ramos, 2006).

A comprehensive characterisation of all *cis*-acting HIV-1 RNA elements is necessary to fully understand the viral gene expression pathway. In order to identify novel putative *cis*-acting HIV-1 RNA elements necessary for viral replication, both Takata *et al.* (Takata et al., 2017) and Antzin-Anduetza *et al.* (Antzin-Anduetza et al., 2017) employed a codon modification approach based around introduction of synonymous mutations into the genomic RNA. Takata *et al.* codon modified the HIV-1 genome in a series of blocks which in total covered a majority of the HIV genome. They identified numerous regions spanning across the genome which were sensitive to codon-modification and able to induce restriction of viral replication. Antzin-Anduetza *et al.* focused on modification of the *gag* region, and observed viral restriction that increased as the length of the codon modified region in *gag* was increased.

Both Takata *et al.* and Antzin-Anduetza *et al.* observed that these codon modifications had inadvertently increased the CpG dinucleotide frequency of the viral genome. The CpG content of many RNA vertebrate viruses is suppressed (Karlin et al., 1994; Rima and McFerran, 1997;

Simmonds et al., 2013), including in HIV-1 (Kypr et al., 1989). This suggests that CpG dinucleotides may be deleterious to RNA viruses and under negative selection. Supporting this, it has been previously shown that increasing the CpG content of picornaviruses and influenza A virus reduces their ability to replicate (Atkinson et al., 2014; Fros et al., 2017; Gaunt et al., 2016; Tulloch et al., 2014).

With these observations, a natural question is thus whether the effects of codon modification on HIV-1 replication are mediated solely through increases in CpG dinucleotide frequency. In light of this, Antzin-Anduetza *et al.* investigated the specific effects of a "CpG-only" codon modification and contrasted this with a codon modification approach that avoided introducing any novel CpGs. This analysis revealed that CpGs were necessary for the observed restriction phenotype, but that modifications to the surrounding nucleotide context present in the fully codon-modified viruses also contributed. In their work, Takata *et al.* used a similar strategy to demonstrate the causal role of CpGs in mediating the codon-modification-induced HIV-1 restriction. Further, Takata *et al.* were able to identify the host protein Zinc anti-viral protein (ZAP) as being necessary for the CpG-mediated restriction. They observed that ZAP directly binds CpGs within the HIV-1 viral RNA, and thus binds to codon modified HIV-1 transcripts with increased frequency relative to wild-type virus. ZAP itself does not possess enzymatic activity. However, Ficarelli *et al.* (Ficarelli et al., 2019) identified KHNYN, a protein with endonuclease activity, as an interacting partner of ZAP that is additionally necessary for the CpG-mediated phenotype. Thus, both ZAP and KHNYN are necessary for CpG-mediated restriction, with KHNYN likely responsible for mediating degradation of ZAP-bound HIV-1 RNA.

6.1.2 ZAP-independent mechanisms of CpG-mediated HIV-1 restriction

We have observed that introduction of CpGs into certain regions of the HIV-1 genome produces a ZAP-independent restriction phenotype. Specifically, codon modification or introduction of CpGs into *gag* resulted in a ZAP-independent reduction in infective virus production, whilst in contrast, codon modification of *env* had ZAP-dependent effects (Figure 6-1). This raises the question of whether an additional antiviral factor may be involved in sensing the CpGs within the *gag* region of these modified viruses. Other mechanisms may be involved. For instance, the modified region of *gag* may contain uncharacterised *cis*-acting elements. Disrupting this region could alter interactions between host proteins and viral RNAs and thus disrupt steps of the gene expression pathway. To investigate these possibilities, we utilized the

codon modified viral constructs from Antzin-Anduetza *et al.* (Antzin-Anduetza et al., 2017) (Figure 6-2, Table 1) to transfect HeLa cells, before extracting RNA for RNA-seq. This allowed investigation of both the host transcriptome and of the expressed HIV-1 transcripts. Since constructs with varying numbers of introduced CpGs were used for transfections, the relative effects of differing CpG loads could be determined.

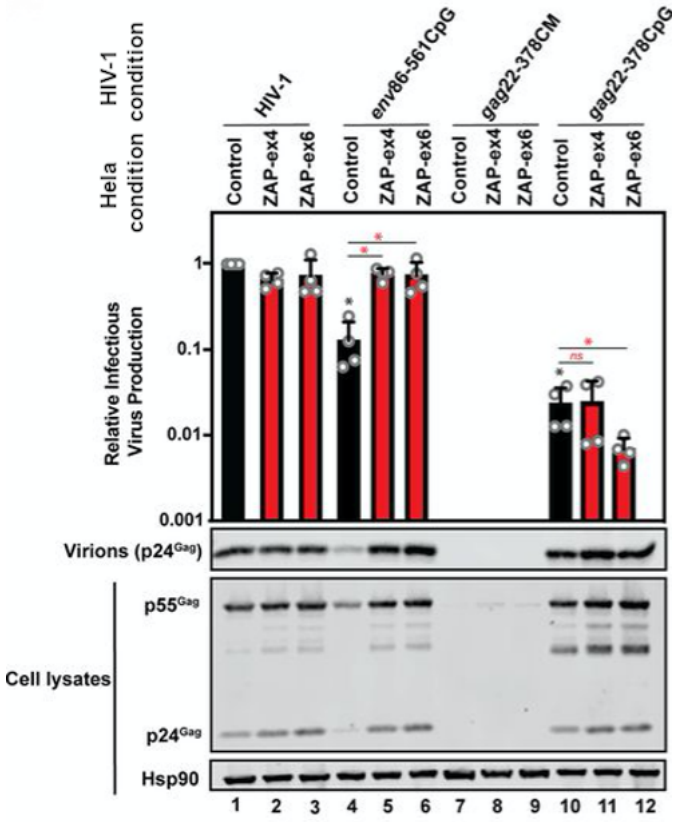

Figure 6-1. Introducing CpG dinucleotides into *gag* has both ZAP-dependent and ZAP-independent effects on HIV-1 replication. Each column shows data from a different experimental condition. Several viral constructs with varying degrees of codon modification were transiently transfected into HeLa cells (Table 1), as depicted via 'HIV-1 condition'. Additionally, three HeLa cell conditions were used, either wild-type control cells, or CRISPR-mediated ZAP knock-out HeLa cells, targeted with one of two guide sequences (Zap-ex4, Zap-ex6), as depicted under 'HeLa condition'. Each row of data then shows the results of a different assay. To measure infectious-virus production, supernatants were collected from these transfected HeLa cells and used to infect TZM-bl reporter cells. Infectivity was determined by induction of β-Galactosidase and subsequent light emission, and normalized to

levels produced upon transfection of control HeLa CRISPR cells with wild-type HIV-1. Top bar charts show the mean of four independent experiments. Error bars show standard deviations. Results of pairwise quantitative comparisons are depicted: * = $P < 0.05$, ns =not significant, as determined by a two-tailed unpaired *t*-test. Red asterisks and ns relate to comparison of Zap CRISPR HeLa cells (red bars) with control CRISPR HeLa cells (black bars). Black asterisks compare control CRISPR HeLa cells transfected with HIV-1, HIV-1$_{gag22–378}$CM, HIV-1$_{gag22–378}$CpG, or HIV-1$_{env88-561}$CpG (black bars). The gel shows results of western blotting for various viral proteins across the different experimental conditions. Expression of the Gag polyprotein p55 and of the Gag product p24 in the cell lysate, and of p24 in the media specifically (Virions), was detected via immunoblotting with Hsp90 as a control. Experimental work performed by Irati Antzin-Anduetza and Mattia Ficarelli (Ficarelli et al., 2020).

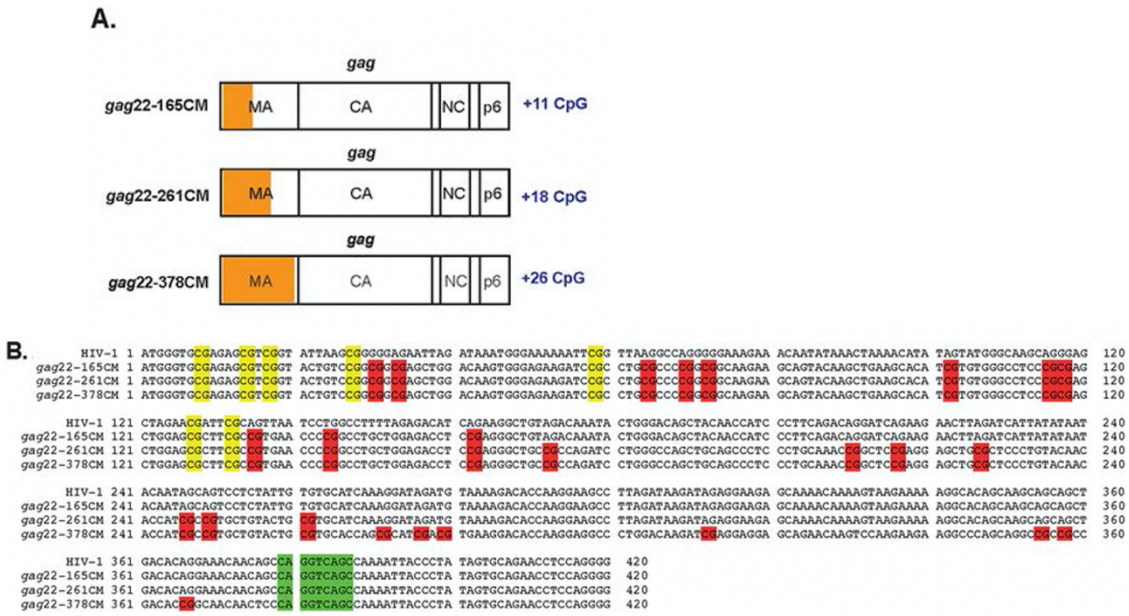

Figure 6-2. Codon modified HIV-1 viral constructs. (A) Schema of the HIV-1 *gag* region with the length of codon modified region depicted in orange for HIV-1$_{gag22–165}$CM, HIV-1$_{gag22–261}$CM, and HIV-1$_{gag22–378}$CM. Numbers of introduced CpG dinucleotides are shown **(B)** Multiple sequence alignment of *gag* regions in HIV-1, HIV-1$_{gag22–165}$CM, HIV-1$_{gag22–261}$CM, and HIV-1$_{gag22–378}$CM. Introduced CpGs are highlighted in red. CpGs found in the wild-type HIV-1 *gag* are highlighted in yellow. Green shows a cryptic splice donor (CD1) that was found to be activated by the synonymous mutations introduced in *gag*, as detailed below (Figure 6-6). CM = codon modification.

Table 6-1. Number of CpGs and mutations introduced by codon modification into HIV-1 constructs. CM = codon modification. Subscript text indicates the region to which mutations were introduced.

Virus	CpGs in wild-type virus	CpGs introduced	Mutations introduced
HIV-1$_{gag22\text{-}165}$CM	4	11	49
HIV-1$_{gag22\text{-}261}$CM	4	18	80
HIV-1$_{gag22\text{-}378}$CM	4	26	109
HIV-1$_{gag22\text{-}378}$CpG	4	26	30
HIV-1$_{env86\text{-}561}$CpG	1	36	43

6.1.3 Aims:

The HIV-1 lifecycle depends upon the interactions between host RBPs and viral transcripts to facilitate processes such as splicing and nuclear export. CpG dinucleotides inhibit HIV-1 replication through the actions of the RBP ZAP, but a ZAP-independent mechanism also appears to exist. We aim to identify the mechanism of this ZAP-independent CpG effect, including through investigating the hypothesis that additional host RBPs with CpG binding preferences may be involved. Specifically, we aim to:

1. Analyse the splicing and abundance of expressed viral transcripts in HeLa cells transfected with codon modified or wild type HIV-1 constructs.
2. Compare host changes in gene expression in response to transfection with codon modified viruses relative to wild type virus.

6.2 Results

6.2.1 Codon modification reduces the abundance of HIV-1 RNA produced in transfected cells

HIV-1 constructs were codon modified to introduce increasing numbers of CpGs into *gag* as detailed in Table 6-1 (and in further detail in Chapter 2 - Methods). HeLa cells were transfected with these constructs using two replicate experiments. For use as control conditions, cells were either left untrasfected or transfected with a viral construct containing the wild-type

unmodified HIV-1 genome. Subsequently, RNA was extracted from lysed cells and used for poly(A) selected RNA-sequencing.

Principal component analysis revealed that after transfection of HeLa cells with either wild type or codon modified HIV-1, the largest source of variation was whether cells were transfected or untransfected (Figure 6-2). Consistent with previous work (Antzin-Anduetza et al., 2017), codon modification reduced the expression of total HIV-1 RNA (Figure 6-3). The magnitude of reduction to HIV-1 RNA abundance increased with successively greater regions of codon modification in *gag*, and results were concordant between RNA-seq (Figure 6-3A) and qPCR (Figure 6-3B).

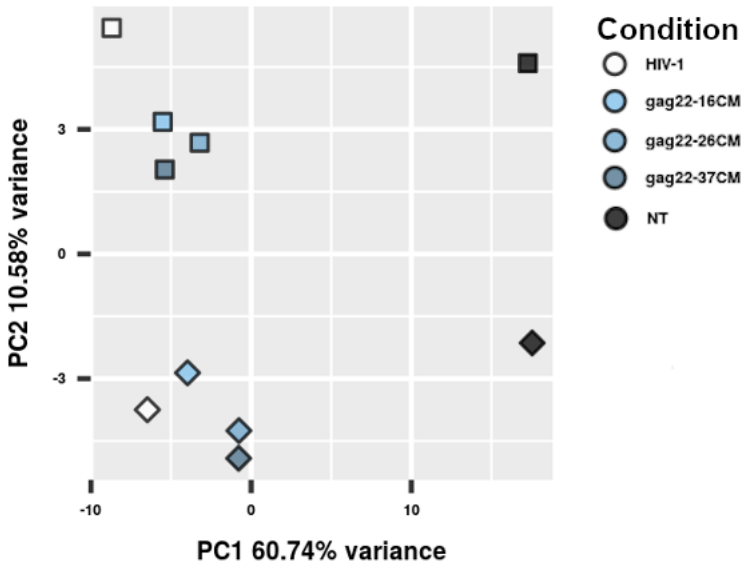

Figure 6-2. Principal component analysis of gene expression after infection with wild-type or codon modified HV-1. Gene expression in HeLa cells after transfection with: wild-type HIV-1, HIV-1$_{gag22–165}$CM, HIV-1$_{gag22–261}$CM, and HIV-1$_{gag22–378}$CM constructs as per Table 6.1. NT = non-transfected control, PC = principal component. Different shapes denote independent replicate experiments. Regularized log transformed TPM values were used as input for PCA.

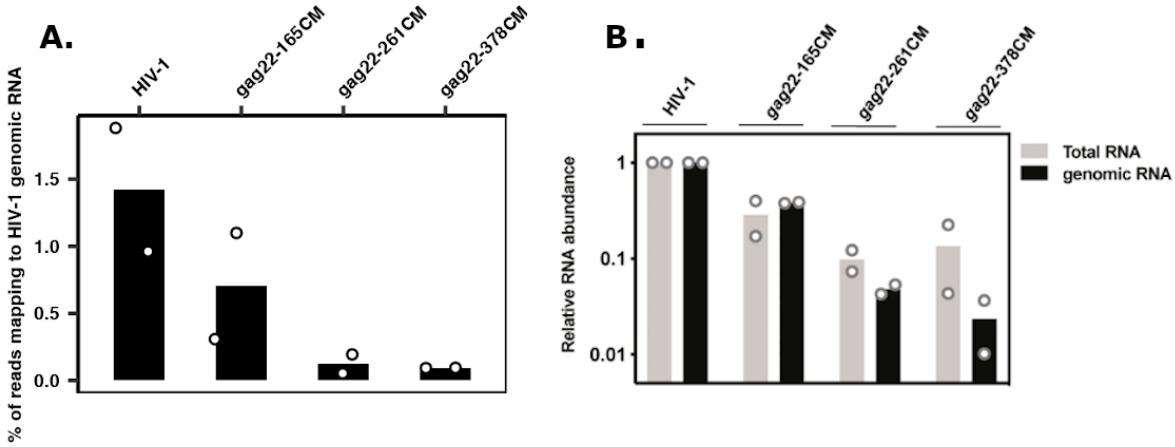

Figure 6-3. Codon modification reduces HIV-1 RNA abundance in transfected HeLa cells. Approximately 48 hours after transfection, HeLa cells were lysed and RNA was extracted for analysis with RNA-seq or qRT-PCR. **(A)** The percentage of total reads in RNA-seq libraries which mapped to the HIV-1 genomic RNA. **(B).** Total and genomic-RNA abundances as quantified by qRT-PCR and normalized to wild-type HIV-1. Genomic RNA is quantified using primers which specifically amplify the full-length unspliced RNA, whilst total RNA is quantified using primers which amplify all HIV-1 transcripts whether spliced or unspliced. Bar charts show the mean of two independent experiments. N.B experimental work in **(B).** performed by Irati Antzin-Anduetza and Mattia Ficarelli.

6.2.2 Codon modification of HIV-1 disrupts splicing of viral transcripts

A possible mechanism by which codon modification could disrupt viral replication is through modulating the splicing of viral transcripts. To investigate HIV-1 alternative splicing, RNA-seq data was used to identify usage of both known and novel splice junctions within HIV-1 transcripts. The relative usage of the identified splice junctions in expressed viral transcripts was then quantified (Figure 6-4). Codon modification resulted in altered patterns of splice site usage, including though an increase in the usage of non-canonical junctions.

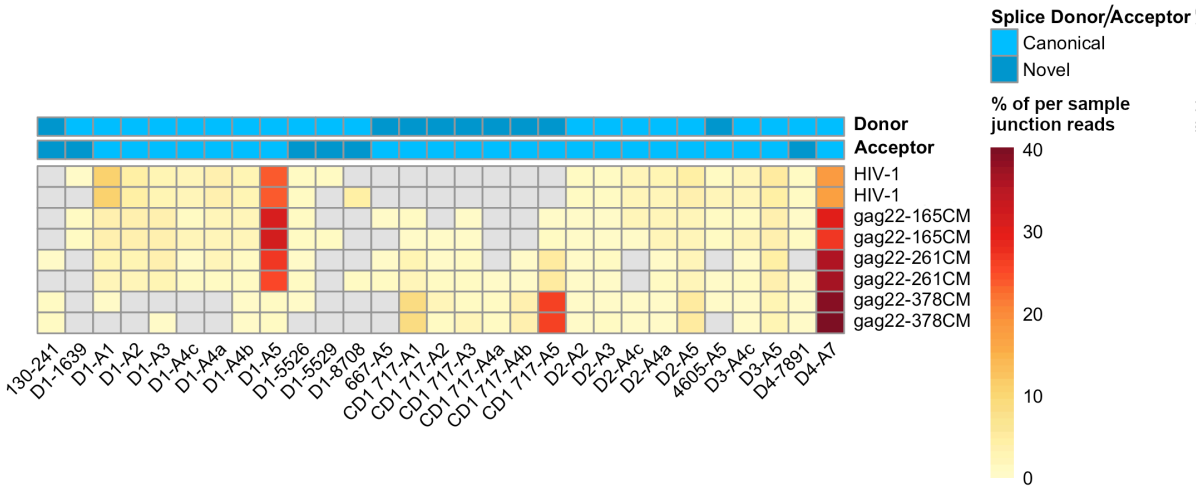

Figure 6-4. Codon modification of HIV-1 induces use of non-canonical splice sites. Each column shows the usage of a unique splice junction. Splice junctions are labelled with the canonical donor (D) and acceptor (A) numbers or the respective genomic RNA co-ordinates for non-canonical splice sites. Rows show pairs of independent replicates of HeLa cells transfected with viral constructs: HIV-$1_{gag22–165}$CM, HIV-$1_{gag22–261}$CM, and HIV-$1_{gag22–378}$CM. Splice junction usage is quantified as the per-sample percentage of junction-spanning HIV-1-mapping RNA-seq reads. Grey boxes indicate that a splice junction was not used in a given sample. Only junctions with relative usage greater than 1% of total HIV-1-mapping junction-spanning reads in at least a single sample are shown. CD = cryptic donor.

The most striking alteration to wild-type HIV-1 splicing was an apparent shift in the usage of canonical donor 1 (D1) for a downstream non-canonical donor (Figure 6-4). Closer examination of this splice event showed an almost perfectly linear shift from the use of D1 to this non-canonical donor with increasing lengths of *gag* codon-optimised sequence (Figure 6-5). Importantly, this donor site is outside of the region of codon modification (Figure 6-2B), and thus is a pre-existing cryptic donor we term cryptic donor 1 (CD1). Activation of the cryptic splice donor acted to increase the length of the first exon incorporated into the spliced viral transcripts to include a region of the *gag* sequence containing the Gag initiation codon (Figure 6-6). Inclusion of this additional start codon could interfere with translation initiation at downstream start codons, including those utilised for translation of proteins encoded by singly or fully spliced mRNAs such as Rev and Tat, which are essential for HIV-1 gene expression. This would decrease viral replication and could account for the observed decrease in viral RNA (Figure 6-3), whilst abundance of gag proteins would remain potentially unaltered (Figure 6-1).

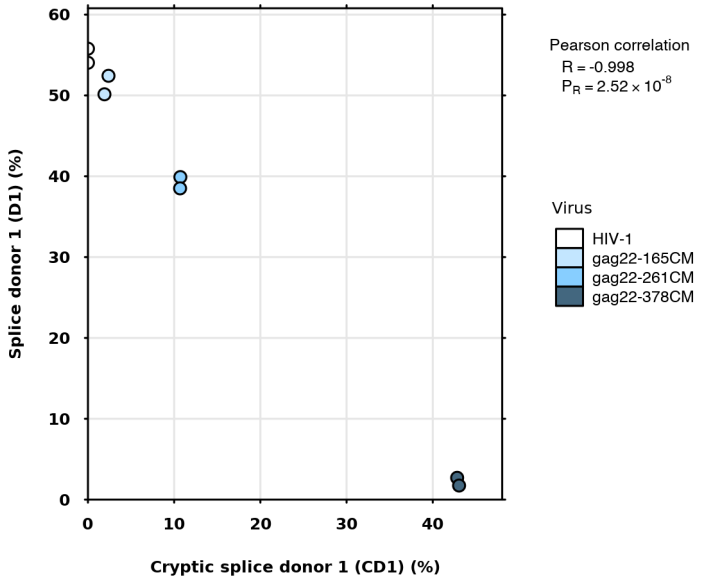

Figure 6-5. Codon modification of *gag* reduces use of canonical splice donor 1 in favour of a cryptic donor. Data shows the per-sample percentage of junction spanning reads using either canonical splice donor 1 or cryptic splice donor 1 with any splice acceptor. Pearson correlation (R) and associated p value (P_R) are shown for the relationship between usage of donor 1 and cryptic donor 1.

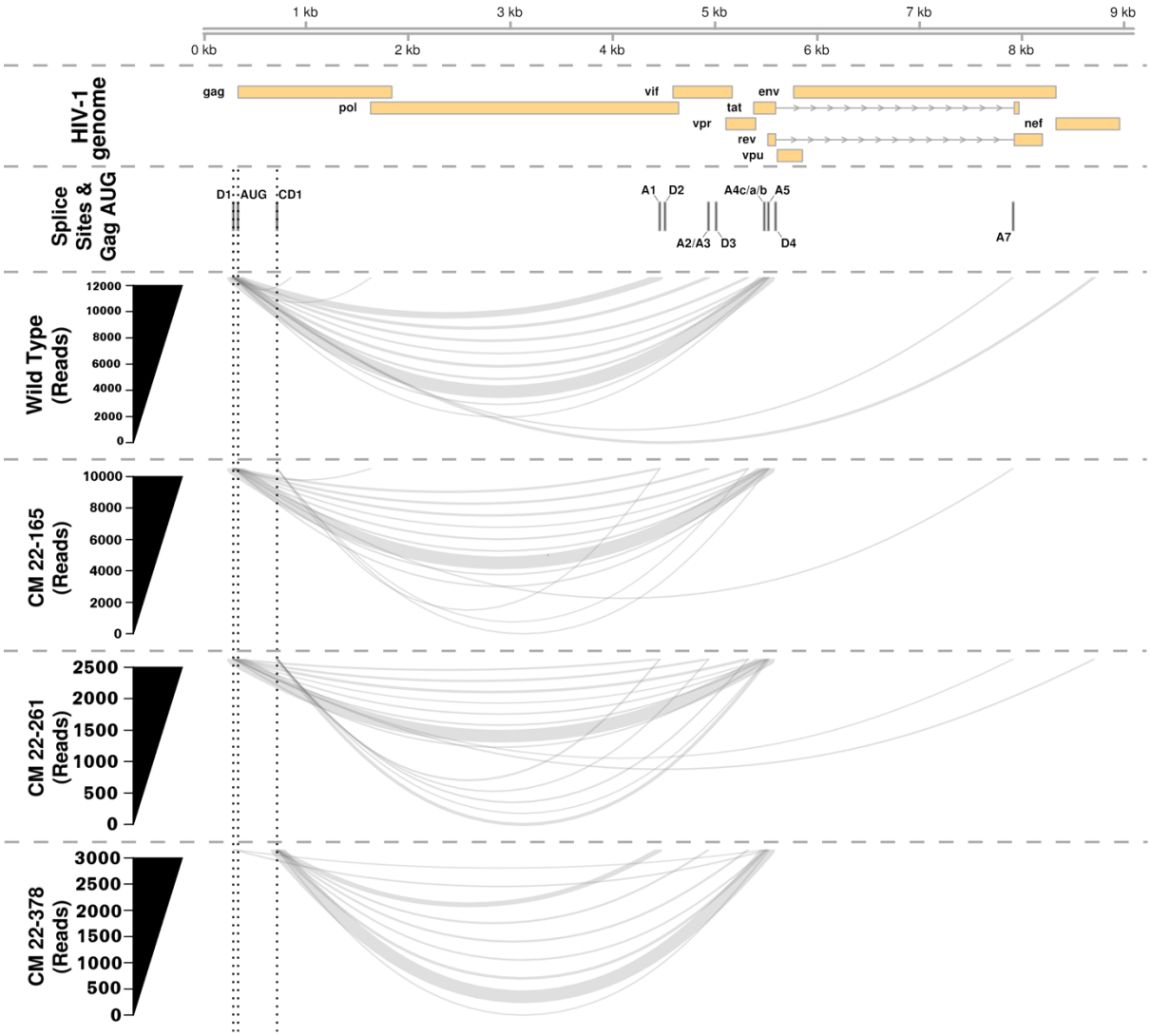

Figure 6-6. Relative usage of splice donor 1 and cryptic donor 1 upon codon modification of HIV-1. The 9173 nt HIV-1 genomic RNA and features are depicted in the "HIV-1 genome" track. The HIV-1 open reading frames are shown as yellow-filled boxes. Canonical donors (D1-4) and acceptors (A1-7), codon modification-induced-cryptic donor (CD1), and gag start codon (AUG), are shown in the "Splice Sites and Gag AUG" track. The numbers of reads supporting use of D1 (nt 290) or CD1 (nt 717) paired with any canonical or non-canonical acceptor and summed across duplicate samples is shown; depicted by line width (y-axis). Line height is arbitrary. Gapped vertical lines trace the position of D1, CD1, and gag AUG.

In light of these findings, we hypothesize that codon modification of *gag* in the region downstream of the cryptic donor may have introduced an exonic splicing enhancer (ESE). Usage of the cryptic donor increased with the number of introduced CpGs and the proximity of

CpGs to the donor site (Figure 6-2B & Figure 6-6). Therefore, a splicing factor which preferentially binds to CpGs may be responsible for the usage of this cryptic donor. Of the RBPs that have well characterised binding preferences, a number have motifs which contain consensus CpG dinucleotides (Figure 6-7) (Ray et al., 2013), and these proteins are thus candidates for splicing enhancers of the cryptic donor. Of course, there are many RBPs with as yet undetermined binding preferences (Gerstberger et al., 2014), and these may also bind to motifs containing CpGs.

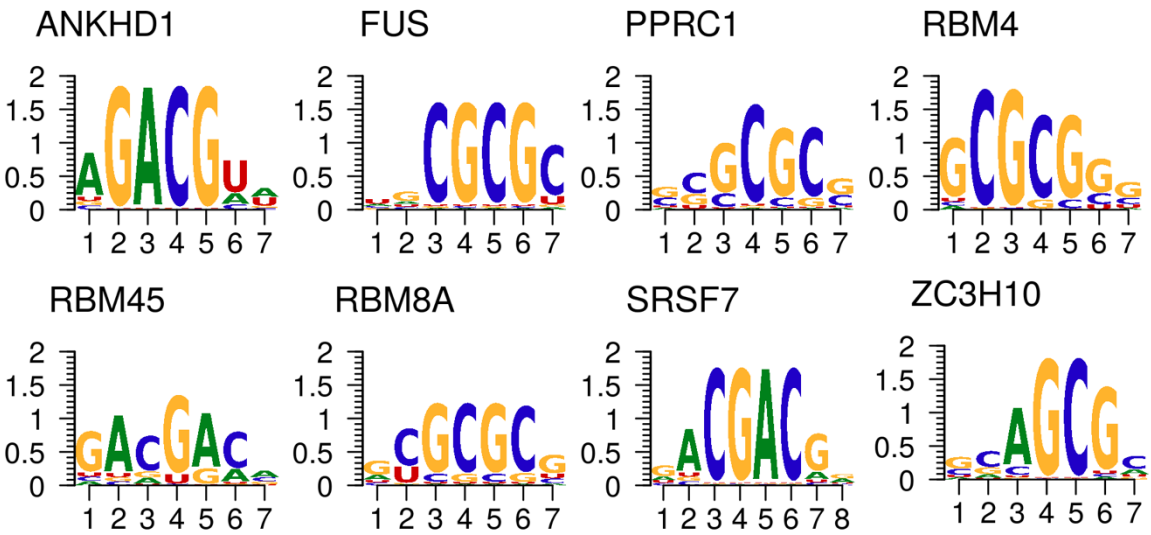

Figure 6-7. RNA-binding proteins with binding motifs containing CpG. Position weight matrices depicted are from (Ray et al., 2013).

6.2.3 Codon modified and wild type HIV-1 induce similar changes in host gene mRNA abundance

The disruption to splicing induced by codon optimization of *gag* may account for the ZAP-independent reductions in infectious viral production (Figure 6-1) and production of viral RNA (Figure 6-3). It is possible that additional mechanisms exist however. For instance, there may be host restriction factors in addition to ZAP which can recognize the codon modified HIV-1 RNA sequences. To identify changes in host gene expression which may reflect activation of an anti-viral response, differential gene expression analysis was performed. Transfection with wild-type HIV-1 induced altered mRNA abundance for 356 genes, relative to untransfected control cells (Figure 6-8A). Transfection with the most highly modified virus (*gag*22-378CM)

produced similar changes in host gene mRNA abundance. Differential expression analysis comparing cells transfected with either wild-type or *gag*22-378CM HIV-1 identified only nine genes with significantly altered mRNA abundance (Figure 6-8B). Eight genes had greater mRNA abundance after transfection with unmodified HIV-1, and a single gene had significantly greater abundance in response to transfection with the codon modified virus. Genes with increased expression after transfection with wild-type HIV-1 may reflect the effects of more productive viral gene expression. For instance, CXCL8 is upregulated upon transfection with wild-type virus (Figure 6-8A) but to a lesser extent after transfection with *gag*22-378CM (Figure 6-8B). Increased CXLC8 expression has been previously shown to promote productive HIV-1 infection (Mamik and Ghorpade, 2014). Increased gene expression in cells transfected with *gag*22-378CM relative to wild-type virus could potentially represent activation of a host anti-viral response. There is a single gene robustly upregulated specifically in *gag*22-378CM transfected cells - AP001267.1 (Figure 6-8B). AP001267.1 is an uncharacterised anti-sense transcript and is an interesting candidate for a ZAP-independent anti-viral gene that recognizes CpGs in the HIV-1 genomic RNA.

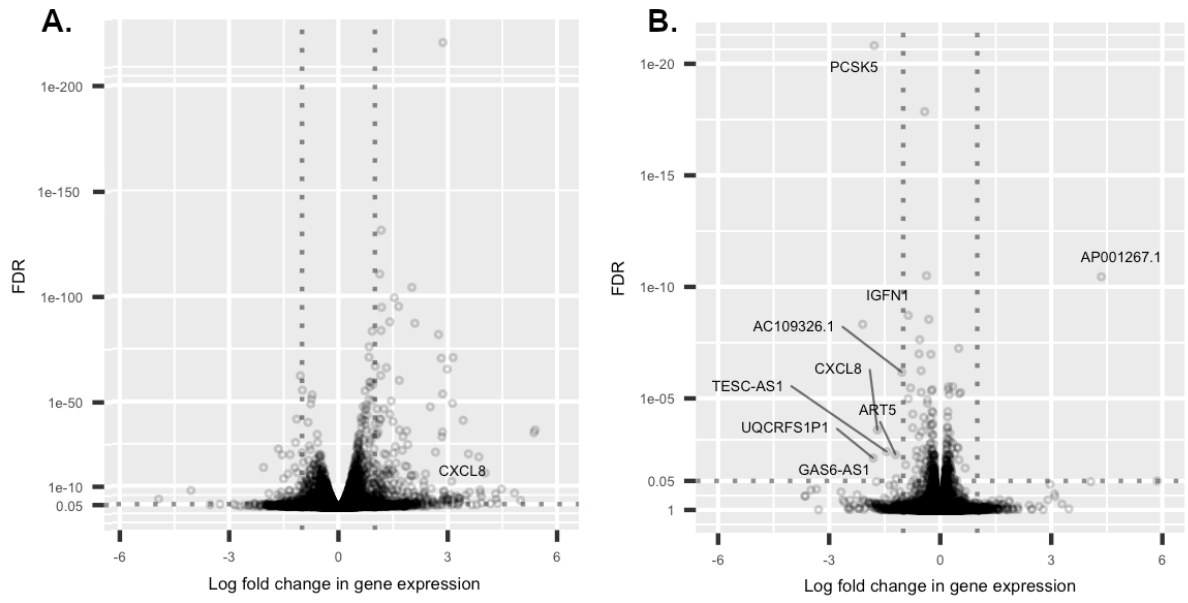

Figure 6-8. Volcano plot of host gene expression upon transfection with wild type and codon modified HIV-1. Gene expression in: **(A)** untransfected control HeLa cells relative to HeLa cells transfected with wild-type HIV-1, **(B)** HeLa cells transfected with *gag*22-378CM HIV-1 relative to wild-type HIV-1. FDR = false discovery rate. Genes considered as differentially expressed

(FDR < 0.05 and log fold change > 1) are labelled. Vertical lines mark log fold change of -1 and 1, whilst the horizontal line marks an FDR of 5%.

6.3 Discussion

Codon modification of the HIV-1 genome has been previously shown to reduce viral replication (Antzin-Anduetza et al., 2017; Takata et al., 2017, 2018). However, the mechanisms by which this occurs have been unclear. Codon modification in the 5' region of gag reduces total and genomic HIV-1 RNA relative to wild-type HIV-1 (Figure 6-3). This phenotype is largely mediated by increased HIV-1 CpG content (Antzin-Anduetza et al., 2017), which sensitizes the virus to both ZAP-dependent and independent effects (Figure 6-2) (Takata et al., 2017). We have shown here that the ZAP-independent CpG-mediated phenotype after *gag* modification appears to be mediated via disrupted splicing due to the activation of a pre-existing cryptic splice donor (Figure 6-5 & 6-6). The usage of this alternative donor is predicted to interfere with subsequent translation of the HIV-1 spliced mRNAs via introduction of the *gag* AUG codon into the first exon of all HIV-1 transcripts (Figure 6-6). This would be predicted to induce inefficient translation of spliced and incompletely spliced HIV-1 mRNAs. Activation of this cryptic donor in response to codon modification of the HIV-1 genome was recently reported independently (Takata et al., 2018), although the role of CpG content in the associated restriction phenotype was not specifically addressed. We hypothesise that a splicing enhancer which binds CpGs may be regulating the use of this cryptic donor. Future work to identify this factor/s could focus on knockdown of splicing factors with known CpG preferences (e.g. those in Figure 6-7) to assess whether the use of the cryptic splice donor in *gag* codon modified HIV-1 transcripts is attenuated.

The observed disruption to viral splicing could account for the totality of the ZAP-independent, codon-modification induced HIV-1 restriction phenotype. However, other mechanisms may also be involved. For instance, a host immune response may be mounted against the codon-modified virus, such as that mediated via ZAP. Through analysis of host gene expression, we identified a long non-coding RNA, AP001267.1, that is more abundant at the mRNA level after transfection with *gag*22-378CM relative to transfection with wild-type HIV-1 (Figure 6-8). This gene could be investigated further, again such as via genetic knockdown, to assess the dependency of the codon-modification associated viral restriction phenotype upon expression of this gene. Further, the introduction of CpGs into the HIV-1 genome could have a number of

additional, currently un-investigated effects, such as through altering local or long-range RNA structures (Lavender et al., 2015), or post-transcriptional modifications such as cytosine methylation (Courtney et al., 2019).

Codon optimisation of viruses has been investigated as a strategy for the creation of vaccines (Gaunt et al., 2016; Tulloch et al., 2014). For instance, infection with a modified influenza A virus with increased CpG frequency protected animals from subsequent lethal challenge to the wild-type virus (Gaunt et al., 2016). Knowledge of the mechanisms underlying such attenuation can facilitate the further design of attenuated viruses. The multiple mechanisms underlying CpG-mediated HIV-1 restriction identified here highlight the complexity in design of an attenuated virus, and this may also be the case for other RNA viruses.

6.4 Conclusion

CpGs are suppressed in the genome of HIV-1 and other RNA viruses, and introduction of CpGs to the HIV-1 genomic RNA has been previously shown to cause ZAP-mediated viral restriction. Here we have identified that introduction of CpGs into *gag* results in aberrant splicing via activation of a cryptic donor, and that this independent of ZAP. Future work could determine whether a known or novel splicing factor with CpG binding preferences regulates the use of this pre-existing cryptic donor.

Chapter 7. Discussion

In this thesis, the over-arching aim was to investigate how alternative pre-mRNA splicing is regulated through the actions of splicing factors mediated through *cis*-acting RNA elements. To this end, I have employed a number of approaches for the analysis of RNA-seq datasets. I have benchmarked the performance characteristics of two different motif-based analysis methods for the identification of regulatory splicing factors. Subsequently, the two approaches were compared further through application to a dataset richer in experimental variables and which represented a potential typical use case for which these methods could be applied. The work detailed in the latter chapters addressed specific biological hypotheses. Firstly, the genome-wide effects of the RBP Sam68 on alternative splicing and mRNA abundance were investigated in primary CD4+ T cells. Additionally, alternative splicing was investigated in the context of HIV-1, a virus with tropism towards CD4 expressing cells. Specifically, the relationship between CpG dinucleotides in the genome of HIV-1 and control of alternative splicing was examined.

7.1 Splicing-based motif enrichment analysis is an effective means to infer regulatory factors

Motif enrichment analysis is a method which involves differential splicing analysis followed by testing for enrichment of splicing factor motifs in RNA sequences flanking the identified differentially spliced events. Whilst our main motivation in using this method was to provide a baseline against which S-MARA could be compared, the formal assessment of splicing factor motif enrichment analysis also has its own merit. To our knowledge, the specificity and sensitivity characteristics of a splicing-focused motif enrichment procedure have not previously been investigated. Indeed, this analysis relied upon the use of an extensive resource of splicing factor knockdowns, such as has been provided only recently through the ENCODE project (Nostrand et al., 2018).

Applied to identifying regulatory motifs associated with RNAi-depleted splicing factors, motif enrichment analysis displayed reasonable sensitivity and specificity as assessed via analysis of the ROC AUC (Figure 4-14). The AUC from the ROC analysis was 0.661, indicating that positive control knockdown splicing factors had more significant enrichment scores than non-knockdown negative control factors ~66% of the time. There are limitations on defining true positives and true negatives in an experimental knockdown system. For instance, unknown

downstream effects may occur such as alterations to activity of other, non-target, splicing factors. Further, the depletion of some splicing factors had low efficiency and, even in the case of efficient knockdown, there may be redundancy between splicing factors acting through similar motifs. However, these results show that splicing factor motif enrichment analyses have merit. The motif enrichment approach used here could be further modified in future work to potentially improve upon the sensitivity and specificity characteristics identified here. Further, this work validates the concept of using RNA motifs to infer regulatory splicing factors through analysis of RNA-seq data.

7.2 S-MARA requires further development

7.2.1 Motif enrichment analysis outperforms S-MARA

In this study, MARA was applied to predict which splicing factors regulate gene expression under specific biological conditions. This is a novel application of this methodology, with MARA most commonly applied to study transcription factors. To adapt MARA, a workflow was implemented to generate a matrix of splice event usage quantifications, and a matched matrix of splicing factor motif counts flanking the RNA sequences of these events (Chapter 3). These two matrices then acted as input to S-MARA. The performance of S-MARA was assessed using data from a large-scale shRNA RBP-knockdown experiment published through the ENCODE resource (Burge et al., 2018; Nostrand et al., 2018). This data provided a powerful resource, containing cells specifically depleted for a range of splicing factors. RNA-sequencing of these cells allowed the investigation of knockdown-induced differential splicing.

Applying MARA to these data allowed quantitative estimates of motif splicing activities. Initial benchmarking analysis showed that MARA could infer splicing factor motif activities that varied in response to knockdown-induced changes in splicing between samples. For a subset of splicing factors that were depleted by RNAi, significant modulation to activity of the associated splicing factor motifs could be identified. However, analysis of the receiver operating characteristics revealed poor sensitivity and specificity properties with regards to identifying motifs associated with splicing factors targeted for depletion (Figure 3-19). This finding contrasts with the respective performance of motif enrichment, which displayed reasonable receiver operating characteristics, as discussed. The relative success of the motif enrichment approach demonstrates that there exists useful information in the distribution of splicing

factor motifs across splice junction regions which can be leveraged to infer splicing factor activities. Thus, the analysis of ENCODE RNAi experiments suggests that S-MARA requires further development to improve its function in towards this aim.

After this initial benchmarking process, S-MARA was applied to a timecourse of CD4+ T cell activation and polarization (Chapter 4). This analysis proved more promising, as S-MARA was able to reveal modules of splicing factor motif activity with distinct profiles suggestive of different splicing regulatory programmes. To identify candidate regulatory motifs, stringent filtering for activity profiles associated with time after TCR stimulation was performed (Figure 4-9). This was combined with additional filtering by gene expression, which led to identification of a set of candidate regulators of splicing during the CD4+ T cell activation process. Application of the hypergeometric test revealed that this gene list was enriched for genes with previously demonstrated roles in the control of alternative splicing during CD4+ T cell activation, in addition to containing novel candidate regulatory splicing factors. Motif enrichment analysis of events differentially spliced after activation, again coupled with filtering of splicing factors with low expression, identified a group of candidate regulatory splicing factors that significantly overlapped with those derived from the S-MARA analysis, albeit not significantly enriched for positive control factors. The more promising results of S-MARA as applied to this timecourse analysis may relate to the increased numbers of replicates in each condition (three), in addition to the biological complexity of the experiment – whereby the timecourse of CD4+ T cell activation and polarisation was sampled with high resolution. Indeed, MARA has primarily been optimised using experiments in which a number of biological conditions were investigated, such as timecourses of development or differentiation (Balwierz et al., 2014; The FANTOM Consortium et al., 2009).

However, as with analysis of the ENCODE knockdown experiments, analysis of the ROC AUC showed an improved performance of the motif enrichment analysis method as compared to S-MARA (Figure 4-15). Indeed, although the AUC for S-MARA was improved relative to that obtained from analysis of the ENCODE data; the 95% confidence intervals still contained the 0.5 value, indicating that the method may not perform greater than chance at specifically identifying motifs associated with positive control splicing factors. Importantly, a potential source of bias in the previously discussed hypergeometric testing was identified. Positive controls had more associated motifs on average as compared to other splicing factors in the

analysis (Figure 4-11). As such, positive control factors were more likely to be identified *ab initio,* and this was not accounted for through hypergeometric testing. However, the ROC analysis was performed directly at the motif level, and was thus able to account for this bias. There are limitations with defining true positive and negative regulatory splicing factors in this analysis. For instance, there are potentially novel unidentified splicing factors which act to control alternative splicing during the CD4+ T cell activation process, and such factors would thus in effect be mis-labelled as negative control cases. However, the results from this ROC AUC analysis are consistent with the analysis of ENCODE shRNA-knockdown data outlined in Chapter 3. Therefore, motif enrichment methods currently appear to outperform S-MARA, which places a caveat over the predictions made through application of S-MARA to the CD4+ T cell activation timecourse.

Our main hypothesis was that S-MARA would perform better than the simpler motif enrichment approach, since it has theoretical advantages. For instance, S-MARA leverages the full genome-wide quantitative splicing information, whilst motif enrichment testing does not make direct use of quantitative splicing information. With this in mind, the greater performance of motif enrichment testing over S-MARA was unexpected. A potential advantage of the motif enrichment procedure is that a reduced set of differentially spliced junctions is used as input, as opposed to the genome-wide splice junction data input to S-MARA. This could provide a stronger signal to noise ratio with regards to motif count features. The central analysis of MARA is a linear regression of quantitative measures of gene activity against motif count features. Several other studies have employed a regression analysis of splicing as a function of sequence features with the aim of identifying regulatory motifs. These studies have all incorporated an initial differential splicing analysis to select only highly differentially spliced events as input to the latter regression stages (Wen et al., 2013; Xin Wang et al., 2008; Zhang et al., 2012). An initial attempt to incorporate a similar strategy was applied here, whereby input to S-MARA was reduced to include only those splice junctions with significantly altered usage after RNAi-mediated splicing factor depletion. However, this approach did not improve the performance of S-MARA in identifying regulatory splicing factors (Figure 3-20).

Another possibility relates to the use of the Wilcoxon rank sum test as the method for inferring motif enrichment. This non-parametric test should be relatively unaffected by characteristics of the underlying motif count distributions for different splicing factor motifs. In contrast, S-

MARA as applied here employs linear modelling of logit transformed PSI values against motif count values. As such, this method assumes a linear relationship between motif count occurrence and logit-PSI values, and could therefore be affected by count distribution features such as the presence of outlier counts.

In this analysis, the presence of potential binding sites was assessed for each RNA sequence of interest using splicing factor PSSMs and a sliding window approach with the RBPmap methodology (Paz et al., 2014). The sum of predicted binding sites per sequence was then used for linear modelling with MARA. In many cases, outlier sequences having hundreds of predicted binding sites/motif counts were observed (e.g. Figure 4-13). In a cell, saturation effects will exist whereby further increases in potential binding sites will not increase the likelihood of splicing factor binding. For instance, an assessment of the binding characteristics of three RBPs (PUM2, QKI, and ELAVL1), revealed that the likelihood of binding increased with the number of high motif match sites in a given sequence up until ten sites were reached, before plateauing or even decreasing with further increases in motif counts (Yu et al., 2019). An initial investigation of potential saturation effects was performed herein, whereby motif counts for a given splice junction region were capped at either 15 or 30. This approach, however, did not alter the performance characteristics of S-MARA (Figure 3-20).

In the application of MARA to transcription factor analysis, various strategies have been applied for estimating transcription factor binding site/motif occurrences. These include using the sum of binding site probabilities across a sequence (Balwierz et al., 2014), a single per-sequence PSSM-match P-value (Madsen et al., 2018), or a sum of the counts of predicted binding sites across a sequence (The FANTOM Consortium et al., 2009). Only the last approach was applied in this study, and there are therefore several alternatives which could potentially improve S-MARA.

7.2.2 Discrepancies between S-MARA and MARA as applied to transcription factor biology

It should be noted that a formal analysis of the ROC AUC of MARA as applied to the analysis of transcription factors has not been performed. However, the predictive power of MARA in identifying positive control regulatory transcription factors has been assessed through other methods (Madsen et al., 2018; The FANTOM Consortium et al., 2009). For instance, the IMAGE

implementation of MARA was able to predict regulators of adipogenesis which showed a greater enrichment for known positive control regulators as compared with simpler analyses of transcription factor gene expression, transcription factor motif enrichment, or a combination of the two (Madsen et al., 2018). Thus, it would appear that S-MARA does not achieve the same performance as "transcription factor-MARA".

A possible source of this discrepancy relates to the underlying properties of RBP motifs, which as a group represent a relatively small subset of the total potential sequence space, and are typically of low entropy (Burge et al., 2018). Indeed, a comparison to transcription factor motifs herein (Figure 3-3) showed that splicing factor motifs were on average lower in information content. Further, previous work in which a deep learning methodology was applied to model both RBP and transcription factor binding preferences found that training RBP models was more challenging, as these models tended to perform worse at predicting both *in vivo and in vitro* true binding affinities (Alipanahi et al., 2015). Analysis of the timecourse of CD4+ T cell activation identified many motifs with highly similar activity profiles, as evidenced by the presence of several large motif modules which were grouped according to their correlative activity profile (Figure 4-4). Unsurprisingly, motifs within a module often shared similar sequence content such as di-nucleotide preferences (Figure 4-5). This highlights the difficulty in resolving activity of motifs with similar PSSMs, which in turn have similar motif count distributions across the genome. Therefore, there may be an inherent challenge to motif-based analyses focused on RBPs.

7.2.3 Limitations and future improvements to S-MARA analysis

7.2.3.1 Importance of RNA secondary structure and higher order sequence features in RBP binding

RNA molecules encode more information than is represented within a PSSM, and this information contributes to RBP binding specificities. Spacing between bound nucleotides, local motif-flanking sequences, and RNA secondary structure, amongst other features, all influence RBP binding specificity (Burge et al., 2018; Taliaferro et al., 2016). A relevant example is the TIA1 and hnRNP C binding motif UUUUU. This motif is represented by a single PSSM in this study, and thus a single motif activity is estimated for both of these splicing factors. However, hnRNP C and TIA1 bind distinct targets *in vivo*, and this is contributed to by structural contexts, with hnRNP C preferentially binding to unstructured RNA relative to TIA1 (Burge et al., 2018).

Importantly, this preference was observed at exons for which hnRNP C binding directly regulates alternative splicing.

Modelling of higher order RNA features has been shown to improve splicing factor binding site predictions for a number splicing factors (Burge et al., 2018; Taliaferro et al., 2016). Alipanahi *et al.* (Alipanahi et al., 2015) developed DeepBind, a deep learning approach for the development of DNA and RNA binding models, and for predicting sequence-specific binding affinities. DeepBind takes as input a set of sequences with associated binding scores derived from any of a number of high-throughput approaches (e.g. HT-SELEX (Jolma et al., 2010) or RNAcompete (Ray et al., 2009)). These sequences are used for *de novo* identification of motifs which are weighted and combined to generate binding models through use of deep convolutional neural networks. DeepBind predictions trained on *in vitro* derived RNAcompete data predicted *in vivo* binding sites (measured through CLIP-seq), with greater accuracy than a simple PSSM-based method. Additionally, DeepBind binding prediction scores showed patterns of enrichment and depletion surrounding splice sites in a manner consistent with roles in splicing regulation. Of note, DeepBind was not compared against the binding site prediction method applied in this study, RBPmap (Paz et al., 2014), which uses PSSMs but additionally takes into account factors such as the propensity of motifs to cluster around true binding sites. Thus, the potential improvement associated with integrating DeepBind into the S-MARA pipeline is unknown and could be further investigated. DeepBind models can be visualised as weighted sets of PSSM-like sequence logos. This visualisation revealed that a variety of sequence features are implicitly captured by the deep learning approach employed, including the presence of sequence position interdependence, variable spacing of bipartite motifs, or the presence of splicing factors with multiple motifs of variable binding affinities. The implicit modelling of these higher levels of sequence complexity likely accounts for the improvement over the more simplistic PSSM-based methods. DeepBind is available as a tool for RBP binding site prediction, with pre-calculated models for the 85 human RBPs studied via RNAcompete by Ray *et al.* (Ray et al., 2013).

The degree to which the DeepBind method implicitly models RNA secondary structures is not reported. Indeed, as recognised by the authors, RNAcompete was not designed to elucidate RBP structural binding preferences (Ray et al., 2009). In 2014, Maticzka *et al.* released GraphProt (Maticzka et al., 2014), a tool specifically developed for the modelling and

predictions of RBP sequence and secondary structural binding preferences. Secondary structures of CLIP-derived RBP-bound RNA sites are predicted *in silico* and utilised in addition to RNA primary sequence information as input to a machine learning based approach for generation of binding models. The resulting models capture nucleotide-specific and position-interdependent structural features, such as the presence of double stranded RNA or hairpin loops, and outperformed previous approaches for binding site prediction (prior to publication of DeepBind which was therefore not included in the comparison). A limitation to these structure-based methods is the non-trivial added computation time and memory needed to estimate complex RNA secondary structures, both during initial model training and subsequent binding prediction steps. Further, GraphProt requires CLIP-seq data for any RBP of interest.

Recently, a new RBP binding site prediction method, beRBP, was published (Yu et al., 2019). The authors employed a hybrid approach, which as per RBPmap, incorporates PSSM-based scoring, analysis of clustering propensity of motif matches, and sequence conservation analysis, but additionally considers RNA sequence accessibility through secondary structure prediction, and consolidates these data into a per-RBP binding model using a Random Forest method. When applied to the prediction of genome-wide *in vivo* bound sequences, as determined through enhanced CLIP (eCLIP), beRBP displayed improved specificity and sensitivity across 26 RBPs as compared to both RBPmap and DeepBind. One limitation to beRBP is that in order to develop RBP-specific binding models, both a PSSM and large numbers of positive and negative sequences, as determined through high-throughput analyses such as RNAcompete, are required as input. This contrasts to RBPmap for which only a PSSM is necessary.

7.2.3.2 Direct Inference of RBP mRNA binding sites

This thesis has focused on the prediction of RBP targets through use of RBP binding models in the form of PSSMs. However, the transcriptome-wide binding targets of individual RBPs can alternatively be directly inferred through the use of immunoprecipitation-sequencing approaches such as CLIP-seq. Indeed, previous analyses have shown improved prediction of splicing regulatory activity using CLIP-based rather than motif based predictions of splicing factor binding (Carazo et al., 2018). In addition to the shRNA-RNA-seq data employed herein, the ENCODE project has generated eCLIP for a range of RBPs in both HepG2 and K562 cell lines

(Van Nostrand et al., 2016). These data would be valuable in future analysis of S-MARA and motif enrichment. For instance, they could be directly incorporated into the work detailed in Chapter 3 to estimate splicing factor binding locations, particularly as the shRNA knockdowns were performed in these same cell lines. The use of CLIP-seq data is thus an additional way in which performance of S-MARA may be improved.

7.2.3.3 Expanding the repertoire of known RBP binding preferences

The above developments to the modelling of RBP binding preferences represent promising strategies through which the accuracy of splicing factor binding site prediction can be improved. Such improvements could help to reduce false negatives of motif-based analyses, in addition to allowing differentiation of binding for splicing factors which share motifs with similar linear motifs (PSSMs). A further avenue for improvement of splicing factor motif-based analyses is to increase the number of splicing factors for which binding models exist. The Attract database (Giudice et al., 2016), which contains the largest single collection of data on RBP-RNA interactions from heterogeneous experimental sources, contains data for approximately only 10% of the ~1500 documented human RBPs (Gerstberger et al., 2014). To date, two large-scale *in vitro* high-throughput analyses of RBP binding preferences have been published, both of which were utilised in this study. Ray *et al.* (Ray et al., 2013) probed 85 human RBPs with RNAcompete, whilst Dominguez *et al.* (Burge et al., 2018) used RNA Bind-n-Seq to study 78 human RBPs. These included 32 and 46 splicing factors respectively, and resulted in motif data for a total of 64 of the custom defined list of 122 splicing factors used in this study. Additionally, Jolma *et al.* (Jolma et al., 2019) recently developed high-throughput RNA-SELEX to capture sequence and structural preferences for 86 human RBPs, although these data are not currently available for public use. High-throughput *in vivo* data has also been utilised to derive RBP motifs. For instance, Feng *et al.* (Feng et al., 2019) utilised eCLIP data for 119 human RBPs produced through the ENCODE project to generate high quality motifs. Future work to expand the number of RBPs with profiled binding preferences will advance both the application of splicing factor activity modelling and the understanding of splicing regulation.

7.2.3.4 Splicing-specific adaptations to the MARA model

The characteristics of our splicing quantification tool of choice, MAJIQ (Vaquero-Garcia et al., 2016), should also be considered. One of the desirable features of MAJIQ is its flexibility with

regards to the types of splicing events that can be profiled. The local splicing variant (LSV) model employed by MAJIQ allows events of arbitrary complexity to be considered, thus capturing the full complexity of alternative splicing across the genome. One limitation to the application of MAJIQ in this study, however, is that different categories of splicing events are not considered separately. That is, a given PSI value is considered in the same manner by S-MARA whether it relates to the relative usage of an exon skipping event, alternative 3' or 5' splice site, intron retention event, or any other splice event type. This method assumes that the regulation of such classes of event does not differ on a global level with regards to the role of splicing factor motifs and corresponding splicing factor activities. Whether this assumption is true is unclear. Deep learning studies aimed at deciphering the splicing code underlying the relative usage of alternative splicing events, and which employed separate models for exon skipping, alternative 3', and alternative 5' splice site usage, have been employed (Busch and Hertel, 2015; Louadi et al., 2019). The relative value of sequence features in predicting usage of these three types of splicing events was assessed, and revealed that the main features discriminating use of each event type related to exon length and the splice donor and acceptor sequences, rather than occurrence of specific *cis*-acting splicing factor motifs. However, S-MARA could be adapted to model the usage of different classes of splicing events separately, so as to investigate the potential effects on improving estimation of splicing factor motif activity. This could be achieved by deconstruction and annotation of LSVs into discrete splice event classes, or through the use of an alternative splicing quantification tool which analyses these events separately *ab initio*, such as VAST-TOOLS (Tapial et al., 2017).

Further improvements to how splicing factor motif activity are modelled could be investigated. As discussed, a limitation of MARA is the assumption of linearity between motif frequencies and regulator activity. This assumption does not account for potential variable repressor-enhancer functions of a given regulator towards different targets. This limitation is pertinent to both splicing factor activity and transcription factor activity, and has been recognised as an area for future development by the MARA developers (Balwierz et al., 2014). Previous approaches to model such dual regulatory roles have been employed. For instance, Bauer *et al*. (Bauer et al., 2010) utilised thermodynamic models to estimate transcription factor activity as a function of motif occurrences and transcription factor binding preferences in *Drosophila melanogaster*. They demonstrated that allowing transcription factor action to vary between repressor and activator on a per transcription unit (*cis*-regulatory module) basis improved

estimates of gene expression. However, this approach was employed to model a limited number of 44 sites, and may need adaptation to be tractable for genome-wide application.

A key difference between the dual activity of splicing factors and transcription factors is the dependence upon splicing factor binding location and the corresponding splicing factor action. For instance, splicing factors can exhibit opposing enhancer or repressor activity when binding a motif located in an exon vs an intron, or when binding upstream or downstream of a regulated exon (Fu and Ares, 2014). As the S-MARA workflow implemented in this study is naïve to such effects, this may negatively affect the performance of S-MARA as compared to the application of MARA to transcription factor biology. One solution includes the use of two input motif count matrices, where the up-and-downstream regions of each splice event or exon are scanned for the presence of motifs independently. Each motif activity would then be modelled twice per sample, with these estimates contrasting for splicing factors with variable activity determined by binding site location. Such an adaptation would be relatively simple to implement as part of future study and may improve splicing factor motif activity estimates. Previous approaches to linear modelling of splice event usage as a function of *in-cis* sequence elements have addressed this aspect of splicing biology by explicitly modelling the regional context of motifs. For instance, Wen *et al.* (Wen et al., 2013) modelled the frequency of putative regulatory sequences in upstream and downstream exonic and intronic regions flanking splicing events as separate coefficients in a lasso regression model. This allowed estimation of splicing factor motif count coefficients (i.e. motif activity) that varied based upon their regional context. A procedure was used to identify potential splicing regulatory elements (SREs) from the set of all possible hexamers, and cassette exon PSI values were then modelled as a function of the frequencies of these hexamers and of potential combinatorial effects between them. This model was able to capture a large percentage of the variance in alternative splicing events across nine human tissues (49.1%–66.5%), performing comparably in this regard to models of gene transcription (Gertz et al., 2009).

7.3 Implications for the understanding of alternative pre-mRNA splicing by RNA-binding proteins

7.3.1 Regulation of alternative splicing during CD4+ T cell activation

A comprehensive understanding of CD4+ T cell function, including the activation of naïve cells in the generation of effector T_h cells, requires elucidation of the full gene expression regulatory networks governing these processes. Re-analysis of the CD4+ T cell activation and polarisation timecourse investigation performed by (Henriksson et al., 2019) using module-based investigation revealed that splicing variation was more strongly associated with the activation process than with CD4+ T cell polarisation. Specifically, CD4+ T cells polarised to a T_{h2} subtype via IL-4 treatment showed broadly similar splicing responses with cells activated through CD3/CD28 stimulation but not exposed to IL-4, a pattern also observed at the transcriptional level by Henriksson *et al*.. The activation stimulus caused splicing modulation to large modules of genes associated with a broad range of biological functions, particularly regarding functions relating to various steps of the gene expression pathway, and suggestive of a contribution to protein production during the activation process (Figure 4-3). The splicing of these genes followed a wide range of temporal patterns of control, from simple switches in splicing behaviour to more complex and intricate modules of splicing regulation (Figure 4-2).

Application of S-MARA combined with linear mixed effect spline modelling identified a set of splicing factor motifs with temporal activity profiles strongly associated time-after activation and consistent across replicates (Figure 4-9). Prioritisation of these motifs and the associated splicing factors provided a method for identifying candidate regulators of the CD4+ T cell gene splicing programme for further study. However, the sensitivity and specificity characteristics of S-MARA needs to be improved and at this time these predictions should not be over-interpreted. Splicing factor motif enrichment was also applied for the inference of putative regulatory motifs and associated factors. Using this approach, a reduced set of motifs showing the strongest association with differential splicing across the activation timecourse was identified (Figure 4-13). This method demonstrated value in identifying regulatory splicing factors (Chapter 3, Figure 3-19). As such, the putative novel regulators of CD4+ T cell activation-mediated differential splicing identified by this method could be prioritized for validation and future investigation.

7.3.2 The role of the RNA-binding protein Sam68 in regulating CD4+ T cell gene expression

The role of the RBP Sam68 in regulating CD4+ T cell gene expression was investigated through an RNAi-based depletion dataset. This analysis allowed identification of Sam68-regulated genes, a subset of which were also regulated in response to activation via CD4+ T cell stimulation (Figure 5-4), and thus potentially represent the contribution of Sam68 to the activation process. The function of these splicing variants, in addition to the specific mechanisms of alternative splicing control could be investigated further, such as by using CLIP-seq based assays to map the Sam68 binding profile across the CD4+ T cell transcriptome. One of the striking observations of this analysis was that Sam68 depletion had a much greater effect on transcript abundance than alternative splicing (Figure 5-8). This finding was somewhat unexpected given the previously documented roles of Sam68 in the regulation of widespread alternative splicing during various cellular developmental processes (Chawla et al., 2009; Huot et al., 2012; Matter et al., 2002). A more detailed investigation into the expression of these genes during the CD4+ T cell activation process could be performed with the data presented in this chapter (Chapter 5). For instance, an investigation into how Sam68 mediates changes in mRNA abundance of these transcripts could be investigated. For example, the upregulation of CD25 upon TCR engagement was shown to depend upon Sam68 binding to the NF-κB complex, which in turn facilitates interaction with the CD25 promoter (Fu et al., 2013). Therefore, a role for NF-κB activity being modulated downstream to Sam68 depletion could be investigated using these data. Further, Sam68 regulates multiple steps of gene expression including transcription, polyadenylation, translation, and splicing (La Rosa et al., 2016; Matter et al., 2002; Pandit et al., 2013; Paronetto et al., 2007, 2009). Therefore, a comprehensive analysis of how Sam68 regulates gene expression should include an investigation into the totality of these RNA processing steps.

7.3.3 The influence of CpG dinucleotides on alternative HIV-1 splicing

HIV-1 undergoes extensive splicing via the host splicing machinery. Here we identified a role for CpG dinucleotides in influencing alternative pre-mRNA splicing of expressed proviral transcripts. Specifically, introducing CpGs into *gag* promoted the use of a cryptic splice donor site that interfered with the use of the canonical donor D1 (Figure 6-6). Interestingly, the number of CpGs introduced correlated with the use of the cryptic donor site (Figure 6-5).

Based on this observation, we hypothesised that a CpG-based splicing enhancer element was introduced into *gag* which promoted the actions of a host splicing factor towards this site. Future work could centre around identification of this hypothesised splicing factor. CpG dinucleotides are suppressed in the HIV-1 genome (Kypr et al., 1989). This appears in at least part to be the result of negative selection, allowing HIV-1 to avoid the effects of the anti-viral protein ZAP, which binds CpG dinucelotides to mediate viral restriction (Ficarelli et al., 2019; Takata et al., 2018). Our findings suggest that CpG suppression also maintains correct HIV-1 alternative splicing. Increasing the CpG frequency of viruses through codon optimisation has been investigated as a potential strategy for viral attenuation in the creation of vaccines (Gaunt et al., 2016). It is important to know the mechanisms of how such codon modification strategies can inhibit viral replication, and we have here shown that splicing is one potential mechanism. Whether CpG suppression in other RNA viruses also influences alternative splicing remains to be investigated.

7.4 Conclusion

In this thesis I have analysed RNA-sequencing datasets to investigate how alternative splicing is regulated through the actions of splicing factors and RNA regulatory elements. To this end, I have employed a novel workflow, S-MARA, for the inference of regulatory splicing factors in a given biological condition. I benchmarked this approach using resources from the ENCODE project. Contrary to expectations, a simple analysis of differential splicing coupled with motif enrichment testing outperformed S-MARA with regards to identification of regulatory splicing factor motifs. When applied to a timecourse of CD4+ T cell activation, candidate splicing regulators including both known regulators and interesting novel candidates were identified through application of both S-MARA and motif enrichment analysis. However, motif enrichment analysis again showed improved sensitivity and specificity characteristics as compared to S-MARA. Therefore, S-MARA currently does not perform favourably. However, improving both the preparation of input data and the MARA model itself will likely enhance performance for the specific aim of identifying regulatory splicing factors. Further, I investigated the genome-wide targets of the RNA binding protein Sam68. This revealed that Sma68 has a more widespread role in regulating mRNA abundance than alternative splicing in a model of CD4+ T cell activation. Finally, a role for CpG suppression in contributing to correct alternative splicing of the HIV-1 genome was identified.

Chapter 8. Appendix

8.1 ENCODE project shRNA RBP knockdown experiment sample accession codes

ENCSR000KYM, ENCSR000YYN, ENCSR003EKR, ENCSR003LSA, ENCSR004OSI, ENCSR004RGI, ENCSR007XKL, ENCSR009PPI, ENCSR010ZMZ, ENCSR011BBS, ENCSR012DAF, ENCSR016IDR, ENCSR016XPB, ENCSR017PRS, ENCSR023HWI, ENCSR024FOF, ENCSR028ITN, ENCSR028YAQ, ENCSR029LGJ, ENCSR030ARO, ENCSR030GZQ, ENCSR031RRO, ENCSR032YMP, ENCSR034VBA, ENCSR040FSN, ENCSR040WAK, ENCSR042QTH, ENCSR047AJA, ENCSR047EEG, ENCSR047IUS, ENCSR047QHX, ENCSR047VPW, ENCSR048BWH, ENCSR052IYH, ENCSR056QEW, ENCSR057GCF, ENCSR060IWW, ENCSR060KRD, ENCSR064DXG, ENCSR066VOO, ENCSR067GHD, ENCSR067LLB, ENCSR070LJO, ENCSR074UZM, ENCSR076PMZ, ENCSR077BPR, ENCSR079IPT, ENCSR079LMZ, ENCSR081IAO, ENCSR081QQH, ENCSR081XRA, ENCSR082UWF, ENCSR082YGI, ENCSR084SCN, ENCSR090UMI, ENCSR092WKG, ENCSR094KBY, ENCSR098NHI, ENCSR101OPF, ENCSR104ABF, ENCSR104OLN, ENCSR105OXX, ENCSR110HAA, ENCSR110ZYD, ENCSR112YTD, ENCSR113PYX, ENCSR116QBU, ENCSR116YMU, ENCSR117WLY, ENCSR118EFE, ENCSR118KUN, ENCSR118VQR, ENCSR118XYK, ENCSR119QWQ, ENCSR124KCF, ENCSR126ARZ, ENCSR129ROE, ENCSR129RWD, ENCSR134JRE, ENCSR135LXL, ENCSR137HKS, ENCSR143COQ, ENCSR143UET, ENCSR147ZBD, ENCSR148MQK, ENCSR149DMY, ENCSR152IWT, ENCSR152MON, ENCSR153GKS, ENCSR154OBA, ENCSR155BMF, ENCSR155EZL, ENCSR164MUK, ENCSR164TLB, ENCSR165BCF, ENCSR165VBD, ENCSR167JPY, ENCSR169QQW, ENCSR174OYC, ENCSR180XTP, ENCSR181RLB, ENCSR182DAW, ENCSR182GKG, ENCSR185JGT, ENCSR188IPO, ENCSR191VWK, ENCSR192BPV, ENCSR192GBD, ENCSR193FFA, ENCSR201WFU, ENCSR205VSQ, ENCSR208GPE, ENCSR210DML, ENCSR210KJB, ENCSR210RWL, ENCSR215FRI, ENCSR219DXZ, ENCSR220TBR, ENCSR222ABK, ENCSR222COT, ENCSR222Csplicing factor, ENCSR222LRL, ENCSR222SMI, ENCSR227AVS, ENCSR230ORC, ENCSR231DXJ, ENCSR232CPD, ENCSR232XRZ, ENCSR234YMW, ENCSR237IWZ, ENCSR237YZT, ENCSR243IGA, ENCSR244SIO, ENCSR245BNJ, ENCSR246RRQ, ENCSR246SOU, ENCSR251ABP, ENCSR253DCB, ENCSR256PLH, ENCSR258VGD, ENCSR264TUE, ENCSR267RHP, ENCSR268JDD, ENCSR269HQA, ENCSR269SJB, ENCSR269ZAO, ENCSR274KWA, ENCSR278CHI, ENCSR279HMU, ENCSR281IUF, ENCSR281KCL, ENCSR286OKW, ENCSR295XKC, ENCSR296ERI, ENCSR300IEW, ENCSR300QFQ, ENCSR302JQA, ENCSR305XWT, ENCSR306EIU, ENCSR306IOF, ENCSR308IKH, ENCSR309HXK, ENCSR309PPC, ENCSR310VND,

ENCSR312SFA, ENCSR312SRB, ENCSR313CHR, ENCSR318HAT, ENCSR318OXM, ENCSR322XVS, ENCSR324WIS, ENCSR325OOM, ENCSR330KHN, ENCSR330YOU, ENCSR334BTA, ENCSR336DFS, ENCSR338CON, ENCSR341PZW, ENCSR342EDG, ENCSR343DHN, ENCSR344XID, ENCSR345VVZ, ENCSR346DZQ, ENCSR347ZHQ, ENCSR354XQY, ENCSR355OQC, ENCSR361LBE, ENCSR362XMY, ENCSR364GRM, ENCSR366FFV, ENCSR372UWV, ENCSR373KOF, ENCSR374NMJ, ENCSR376FGR, ENCSR376RJN, ENCSR379VXW, ENCSR382Qknockdown, ENCSR384BDV, ENCSR385KOY, ENCSR385TMY, ENCSR385UPQ, ENCSR386YEV, ENCSR388CNS, ENCSR389HFU, ENCSR392HSJ, ENCSR395FYF, ENCSR398GHW, ENCSR398HXV, ENCSR398LZW, ENCSR408SDL, ENCSR409CSO, ENCSR410UHJ, ENCSR410ZPU, ENCSR416ZJH, ENCSR419JMU, ENCSR422JMS, ENCSR424JSU, ENCSR424QCW, ENCSR424YSV, ENCSR426UUG, ENCSR438MDN, ENCSR448JAM, ENCSR450VQO, ENCSR453HKS, ENCSR454KYR, ENCSR455VZH, ENCSR457WBK, ENCSR459EMR, ENCSR464ADT, ENCSR471GIS, ENCSR477TRX, ENCSR478FJK, ENCSR481AYC, ENCSR485ZTC, ENCSR486AIO, ENCSR490DYI, ENCSR491FOC, ENCSR492BKM, ENCSR492UFS, ENCSR494UDF, ENCSR494VSD, ENCSR496ETJ, ENCSR500WHE, ENCSR509LIV, ENCSR511BNY, ENCSR511SYK, ENCSR517JDK, ENCSR517JHY, ENCSR518JXY, ENCSR519KXM, ENCSR524YXQ, ENCSR527IVX, ENCSR527QNC, ENCSR528ASX, ENCSR529JNJ, ENCSR529MBZ, ENCSR529QEZ, ENCSR530BOP, ENCSR532ZPP, ENCSR533HXS, ENCSR535YPK, ENCSR538QOG, ENCSR542ESY, ENCSR545AIK, ENCSR546MBH, ENCSR547NWD, ENCSR552NBS, ENCSR555LCE, ENCSR556FNN, ENCSR558XNA, ENCSR560AYQ, ENCSR560RSZ, ENCSR561CBC, ENCSR562CCA, ENCSR563YIS, ENCSR570CWH, ENCSR572AMC, ENCSR572FFX, ENCSR573UBF, ENCSR576GOW, ENCSR577OVP, ENCSR577XBW, ENCSR584JRB, ENCSR584LDM, ENCSR584UYK, ENCSR585KOJ, ENCSR594DNW, ENCSR597IYB, ENCSR597XHH, ENCSR598GKQ, ENCSR598YQX, ENCSR599PXD, ENCSR599UDS, ENCSR602AWR, ENCSR603TCV, ENCSR605MFS, ENCSR606QIX, ENCSR608IAI, ENCSR610AEI, ENCSR610VTA, ENCSR611LQB, ENCSR611ZAL, ENCSR618IQH, ENCSR620HAA, ENCSR620OKS, ENCSR620PUP, ENCSR624FBY, ENCSR624OUI, ENCSR624XHG, ENCSR629EWX, ENCSR629RUG, ENCSR631RFX, ENCSR634KBO, ENCSR634KHL, ENCSR635BOO, ENCSR635FRH, ENCSR637JLM, ENCSR639LKS, ENCSR643UFV, ENCSR644AIM, ENCSR647NYX, ENCSR648BSC, ENCSR648QFY, ENCSR656DQV, ENCSR660ETT, ENCSR660MZN, ENCSR661HEL, ENCSR667PLJ, ENCSR667RIA, ENCSR674knockdownQ, ENCSR674KEK, ENCSR676EKU, ENCSR678MVE, ENCSR678WOA, ENCSR681SMT, ENCSR684HTV, ENCSR685JXU, ENCSR688GVV, ENCSR689MIY, ENCSR689PHN, ENCSR689ZJC, ENCSR691IVR, ENCSR693MZJ, ENCSR694LKY, ENCSR695XOD, ENCSR696JWA, ENCSR696LLZ, ENCSR701GSV, ENCSR706SXN, ENCSR708GKW, ENCSR710NWE, ENCSR711ZJQ, ENCSR712CSN, ENCSR713OLV, ENCSR715XZS, ENCSR716WZH, ENCSR717SJA, ENCSR718EWL,

ENCSR720BPO, ENCSR721MXZ, ENCSR728BOL, ENCSR732IYM, ENCSR741YCA, ENCSR744PAQ, ENCSR744YVR, ENCSR745WVZ, ENCSR746EKS, ENCSR746NIM, ENCSR754RJA, ENCSR755KOM, ENCSR760EGM, ENCSR762FEO, ENCSR767LLP, ENCSR769GES, ENCSR770LYW, ENCSR770OWW, ENCSR771QMJ, ENCSR774BXV, ENCSR775TMW, ENCSR776SXA, ENCSR777EDL, ENCSR778AJO, ENCSR778RWJ, ENCSR778SIU, ENCSR778WPL, ENCSR780YFF, ENCSR781YNI, ENCSR782MXN, ENCSR783LUA, ENCSR783YSQ, ENCSR784FTX, ENCSR788HVK, ENCSR788YGG, ENCSR792CBM, ENCSR792XFP, ENCSR794NUE, ENCSR795VAK, ENCSR807ODB, ENCSR808FBR, ENCSR809ISU, ENCSR810FHY, ENCSR810JYX, ENCSR812EIA, ENCSR812TLY, ENCSR813NZP, ENCSR815CVQ, ENCSR815JDY, ENCSR818TZM, ENCSR820ROH, ENCSR823WTA, ENCSR825QXH, ENCSR831YGP, ENCSR835RMN, ENCSR837QDN, ENCSR838SMC, ENCSR840QOH, ENCSR843LYF, ENCSR844QNT, ENCSR849STR, ENCSR850CKU, ENCSR850FEH, ENCSR850PWM, ENCSR851KEX, ENCSR853PBF, ENCSR853ZJS, ENCSR856CJK, ENCSR856ZRV, ENCSR861ENA, ENCSR866XLI, ENCSR871BXO, ENCSR874DVZ, ENCSR874ZLI, ENCSR880DEH, ENCSR883BXR, ENCSR885YOI, ENCSR891AXF, ENCSR891DYO, ENCSR896CFV, ENCSR896MMU, ENCSR898OPN, ENCSR902WSK, ENCSR904BCZ, ENCSR904CJQ, ENCSR905HID, ENCSR906RHU, ENCSR906WTM, ENCSR907UTB, ENCSR910ECL, ENCSR910YNJ, ENCSR911DGK, ENCSR913CAE, ENCSR913ZWR, ENCSR914WQV, ENCSR916WOI, ENCSR921knockdownS, ENCSR925RNE, ENCSR925SYZ, ENCSR927JXU, ENCSR927SLP, ENCSR927TSP, ENCSR927XBT, ENCSR929PXS, ENCSR939ZRA, ENCSR942MBU, ENCSR943LIB, ENCSR945GUR, ENCSR945UYL, ENCSR945XKW, ENCSR946OFN, ENCSR947OIM, ENCSR952OOP, ENCSR952QDQ, ENCSR952RRH, ENCSR954HAY, ENCSR957EEG, ENCSR958KSY, ENCSR958NDU, ENCSR960MSV, ENCSR961WVL, ENCSR961YAG, ENCSR963RLK, ENCSR964YTW, ENCSR967QNT, ENCSR968BBQ, ENCSR968YWY, ENCSR973QSV, ENCSR978CSQ, ENCSR984CLJ, ENCSR992JGE, ENCSR995JMS, ENCSR995RPB, ENCSR995ZGJ, ENCSR997FOT, ENCSR997HCQ, ENCSR998MZP, ENCSR998RZI

8.2 Henriksson et al. 2019 CD4+ T cell activation and polarisation timecourse RNA-seq sample accessions

ERX2271389, ERX2271390, ERX2271391, ERX2271392, ERX2271393, ERX2271394, ERX2271395, ERX2271396, ERX2271397, ERX2271398, ERX2271399, ERX2271400, ERX2271401, ERX2271402, ERX2271403, ERX2271404, ERX2271405, ERX2271406, ERX2271407, ERX2271408, ERX2271409, ERX2271410, ERX2271411, ERX2271412, ERX2271413, ERX2271414, ERX2271415, ERX2271416, ERX2271417, ERX2271418, ERX2271419, ERX2271420, ERX2271421, ERX2271422, ERX2271423, ERX2271424,

ERX2271425, ERX2271426, ERX2271427, ERX2271428, ERX2271429, ERX2271430, ERX2271431, ERX2271432, ERX2271433, ERX2271434, ERX2271435, ERX2271436, ERX2271437, ERX2271438, ERX2271439, ERX2271440, ERX2271441, ERX2271442, ERX2271443, ERX2271444, ERX2271445

8.3 Splicing factors analysed in this study

Splicing Factor	Does the splicing factor have a PSSM which was used in this study?
CELF1	TRUE
CELF2	FALSE
CELF3	FALSE
CELF4	TRUE
CELF5	TRUE
CELF6	TRUE
DAZAP1	TRUE
ELAVL1	TRUE
ELAVL2	FALSE
ESRP1	TRUE
ESRP2	TRUE
FMR1	TRUE
FUS	TRUE
HNRNPA0	TRUE
HNRNPA1	TRUE
HNRNPA1L2	TRUE
HNRNPA2B1	TRUE
HNRNPA3	FALSE
HNRNPC	TRUE
HNRNPD	TRUE
HNRNPF	TRUE
HNRNPH1	TRUE
HNRNPH2	TRUE
HNRNPH3	FALSE

HNRNPK	TRUE
HNRNPL	TRUE
HNRNPLL	TRUE
HNRNPM	TRUE
HNRNPR	FALSE
HNRNPU	TRUE
HNRNPUL1	FALSE
KHDRBS1	TRUE
KHDRBS2	TRUE
KHDRBS3	TRUE
KHSRP	TRUE
MBNL1	TRUE
MBNL2	FALSE
MBNL3	FALSE
NOVA1	TRUE
NOVA2	FALSE
PCBP1	TRUE
PCBP2	TRUE
PTBP1	TRUE
PTBP2	FALSE
QKI	TRUE
RALY	TRUE
RBFOX1	TRUE
RBFOX2	TRUE
RBFOX3	TRUE
RBM10	FALSE
RBM11	FALSE
RBM15	FALSE
RBM15B	TRUE
RBM17	FALSE
RBM20	FALSE
RBM22	TRUE

RBM24	TRUE
RBM25	TRUE
RBM3	TRUE
RBM38	TRUE
RBM39	FALSE
RBM4	TRUE
RBM41	TRUE
RBM4B	TRUE
RBM5	TRUE
RBM7	FALSE
RBM8A	TRUE
RBMX	FALSE
RBMXL1	FALSE
RBMY1A1	FALSE
RBMY1B	FALSE
RBMY1D	FALSE
RBMY1E	FALSE
RBMY1F	FALSE
RBMY1J	FALSE
SF1	TRUE
SF3B1	FALSE
SFPQ	TRUE
SRPK2	FALSE
SRRM1	FALSE
SRRM2	FALSE
SRRM4	FALSE
SRSF1	TRUE
SRSF10	TRUE
SRSF11	TRUE
SRSF12	FALSE
SRSF2	TRUE
SRSF3	TRUE

SRSF4	TRUE
SRSF5	TRUE
SRSF6	TRUE
SRSF7	TRUE
SRSF8	TRUE
SRSF9	TRUE
SYNCRIP	FALSE
TARDBP	TRUE
TIA1	TRUE
TRA2A	TRUE
TRA2B	TRUE
U2AF1	FALSE
U2AF1L4	FALSE
U2AF2	TRUE
YBX1	TRUE
ZNF638	TRUE
ZRANB2	FALSE
ZRSR2	FALSE
DDX17	FALSE
DDX5	FALSE
ELAVL3	FALSE
ELAVL4	TRUE
HNRNPAB	FALSE
HNRNPCL1	TRUE
HNRNPDL	TRUE
HNRNPUL2	FALSE
MATR3	TRUE
TIAL1	FALSE
SRRM3	FALSE
DHX8	FALSE
HTATSF1	FALSE
LUC7L	FALSE

LUC7L3	FALSE
EWSR1	TRUE

8.4 Lists of genes with regulated alternative splicing in the Sam68 knock down experiments – Chapter 5

Genes with altered splicing upon Sam68 knockdown in resting CD4+ T cells

AGAP4, AGO3, ANKRD36, CASP8, CIC, CLDND1, COPS8, CSRNP1, DCAF8, LRRN3, MBD5, MBNL1, MELK, METTL21A, NOL12, OSBPL1A, PCGF1, PMM2, PPP1R12B, RP11-1035H13.3, RP11-156P1.3, RP11-206L10.2, RP11-223C24.1, RP1-178F15.5, SATB1-AS1, TBC1D3D, TCFL5, TYW5, UBXN8, ZNF829

Genes with altered splicing upon sam68 knockdown in activated CD4+ T cells

AGO3, BICDL1, BTBD1, C5orf63, CASP8, CCR6, CTC-326K19.6, DDX5, GALM, GOLGA8B, GPAT2P1, IL4I1, ITIH4, JPX, LDLRAD4, LRIF1, LST1, PCMTD1, PER2, PREPL, RBM41, RP11-223C24.1, RP11-681B3.4, RP1-178F15.5, RRP7BP, SNX5, SYCP2, TMEM116

Genes with altered splicing upon activation in CD4+ T cells

ABC7-42389800N19.1, ABCA5, ABCC4, ABL1, AC006129.2, AC093616.4, AC147651.4, ACIN1, ADAP1, ADAT2, AGO3, AHRR, AKAP13, AKAP2, ALG13, ALKBH6, ANK3, ANKRD13D, ANKRD28, ANKRD49, APOO, APP, ARGLU1, ARHGAP17, ARRDC2, ASB2, ASXL1, ATAD3B, ATE1, ATP8A1, ATXN7, BACH2, BCL2, BCL2L1, BDH1, BOD1, BORA, BPTF, BSCL2, BTAF1, BTN2A2, C19orf12, C1orf228, C1orf52, C22orf29, C2orf48, C9orf72, CABIN1, CALM1, CARS2, CASK, CASP8, CBX3, CCDC82, CCM2, CCNB1IP1, CCND3, CCNL2, CCNT1, CDK6, CELF2, CEP128, CEP63, CFLAR, CH17-264B6.3, CH17-472G23.4, CHORDC1, CIITA, CLASP1, CLASRP, CMAHP, CNOT1, CNTRL, COA1, COG1, COPS7B, COQ7, CPNE7, CREB3L2, CREBZF, CRELD1, CREM, CSGALNACT1, CTC-459F4.3, CUTA, DBP, DGKD, DGKE, DHX30, DHX34, DIDO1, DNAJC25, DPH7, DTWD1, DYNC1LI2, E2F8, ECE1, EFCAB14, EGLN3, EIF2AK1, EIF4G2, ELF4, EML2, ENGASE, ENO3, ENTPD1-AS1, EP400, EPB41, EPC1, ESPL1, EXOC7, EXOG, FAM107B, FAM122B, FAM136A, FAM221A, FAM49B, FAM60A, FAM65B, FBLN5, FBRSL1, FGFR1, FLVCR1, FOXN3, FOXP1, FPGS, FRG1BP, G2E3, GALNS, GK, GK5, GLUD1P3, GLUL, GOLGA4, GOLGA8B, GON4L, GORAB, GOSR2, GPBP1L1, GPR137, GPR180, GPR19, GUF1, H2AFY2, HDAC4, HDLBP, HEATR3, HERC2P9, HEXIM2, HIBCH,

HMCES, HMGB1, HNRNPH1, HS3ST3B1, HYAL3, IER3, IFT27, IL18BP, IMMP2L, INCENP, INF2, INTS6L, IQCG, IRF4, KDM2A, KDM2B, KDM6B, KIAA0391, KIAA1671, KIF1B, KIF21A, KLC1, KLHDC4, KLHL24, KLHL7, L3HYPDH, LIMS1, LINC-PINT, LRRC23, LRRC27, LRRFIP2, LTA, LTBP3, LUC7L, LUC7L2, LUC7L3, LYAR, MACF1, MAD1L1, MAGED2, MAP2K7, MAP3K1, MAPK8, MAPK9, MBNL1, MBOAT1, MCM7, METTL3, MFSD13A, MGA, MIB2, MICA, MINDY3, MLLT10, MMP24-AS1, MPRIP, MPV17, MRI1, MRPL32, MSTO2P, MTERF2, MTMR4, MTRF1, MYEF2, N6AMT1, NAA38, NABP1, NADK, NAP1L4, NAPB, NBPF9, NCOA2, NCOA5, NCOA7, NDUFAF5, NDUFAF6, NDUFAF7, NEDD9, NET1, NFATC1, NFE2L2, NFIC, NKTR, NOC2L, NPEPL1, NPIPB14P, NPIPB5, NR2C1, NR3C1, NSD3, NT5C3A, NUMA1, NUMB, NUP43, NXT2, OGFOD2, ORAOV1, P2RX4, PACRGL, PAM16, PARGP1, PARP12, PARP8, PASK, PAXBP1, PCGF3, PCMTD2, PCNX2, PDCD6, PER1, PFAS, PHF19, PIDD1, PIGB, PIGL, PISD, PITPNM2, PKD1P5, Pknockdown1P6, PLCXD1, PMM2, PMS2P7, PNRC1, POC1B-AS1, POMT1, POU2F2, PPFIA1, PPIEL, PPM1K, PPP2R5C, PPP3CB, PPP4R3A, PRAG1, PRR3, PSMA1, PTK2B, PTPN4, PVT1, PXN, QRICH1, RABL2A, RALGAPA1, RALGAPB, RALGDS, RBM19, RBM25, RBM28, RBM39, RBM41, RC3H1, RCAN1, RCC1, RCOR3, REPIN1, RERE, REV3L, RGS19, RGS3, RNASEH1-AS1, RNASEL, RNFT1, RNPC3, RP11-1035H13.3, RP11-206L10.2, RP11-43F13.1, RP11-465B22.3, RP11-479O9.4, RP11-514P8.7, RP1-178F15.5, RP13-942N8.1, RPL6, RPP25L, RPS18, RPS9, RPTOR, RREB1, RRP7BP, RSRP1, RTN4, RWDD2A, S100A13, SATB1, SBDSP1, SBNO2, SCAI, SCLT1, SDHAP3, SEC61A2, SELENOI, SEMA4D, SEPT9, SFMBT1, SFMBT2, SFT2D1, SGK1, SH3YL1, SIN3B, SIPA1, SKP2, SLC25A13, SLC25A14, SLC25A30, SLC2A8, SLC38A1, SLC39A10, SLMAP, SMARCD1, SMC5, SMG1P2, SMN2, SNHG4, SNHG7, SNRNP27, SNRNP70, SNRPA1, SOCS2, SPATA33, SPG20, SRBD1, SREBF1, SREK1, SRPK2, SRRM1, SRRT, SRSF10, SRSF11, SRSF2, SS18L1, ST20, ST6GALNAC6, STIM2, STX2, SUPT3H, SVIL-AS1, TAOK3, TARBP1, TARBP2, TBC1D1, TBC1D4, TBCCD1, TBCD, TBRG1, TCAF2, TEPSIN, THAP4, THAP6, THOC3, THUMPD3-AS1, TIA1, TIAL1, TJP2, TM2D2, TMEM116, TMEM126B, TMEM156, TMEM187, TMEM62, TMEM63B, TMEM91, TMTC4, TNFAIP8, TNFRSF25, TNKS1BP1, TNPO2, TOGARAM1, TP53BP1, TP53I13, TRA2A, TRABD, TRAF2, TRAF3IP3, TRERF1, TRIB1, TRIM14, TRIM5, TRMU, TRNT1, TTC12, TTC13, TTC32, TUG1, UBA1, UBE2V1, UBXN8, UNKL, USP9Y, UTY, UXS1, VEZT, WASH2P, WDR4, WRB, XPNPEP3, YIPF1, ZBTB41, ZBTB7B, ZFP57, ZMAT5, ZMIZ1, ZMIZ2, ZMYM6, ZNF213-AS1, ZNF28, ZNF326, ZNF518B, ZNF580, ZNF708, ZNF738, ZNF765, ZNF783, ZNF83, ZRANB2, ZZZ3

Reference List

Adachi, A., Gendelman, H.E., Koenig, S., Folks, T., Willey, R., Rabson, A., and Martin, M.A. (1986). Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. J. Virol. *59*, 284–291.

Afkarian, M., Sedy, J.R., Yang, J., Jacobson, N.G., Cereb, N., Yang, S.Y., Murphy, T.L., and Murphy, K.M. (2002). T-bet is a STAT1-induced regulator of IL-12R expression in naïve CD4+ T cells. Nat. Immunol. *3*, 549–557.

Aghamirzaie, D., Collakova, E., Li, S., and Grene, R. (2016). CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. BMC Genomics *17*, 845.

Akira, S., Uematsu, S., and Takeuchi, O. (2006). Pathogen Recognition and Innate Immunity. Cell *124*, 783–801.

Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA *21*, 1521–1531.

Alexander, D.R. (2000). The CD45 tyrosine phosphatase: a positive and negative regulator of immune cell function. Semin. Immunol. *12*, 349–359.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Alkhatib, G., Combadiere, C., Broder, C.C., Feng, Y., Kennedy, P.E., Murphy, P.M., and Berger, E.A. (1996). CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science *272*, 1955–1958.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol. *13*, R17.

Antzin-Anduetza, I., Mahiet, C., Granger, L.A., Odendall, C., and Swanson, C.M. (2017). Increasing the CpG dinucleotide abundance in the HIV-1 genomic RNA inhibits viral replication. Retrovirology *14*.

Arnold, P., Erb, I., Pachkov, M., Molina, N., and van Nimwegen, E. (2012). MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinforma. Oxf. Engl. *28*, 487–494.

Asmal, M., Colgan, J., Naef, F., Yu, B., Lee, Y., Magnasco, M., and Luban, J. (2003). Production of Ribosome Components in Effector CD4+ T Cells Is Accelerated by TCR Stimulation and Coordinated by ERK-MAPK. Immunity *19*, 535–548.

Atkinson, N.J., Witteveldt, J., Evans, D.J., and Simmonds, P. (2014). The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate

cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Res. *42*, 4527–4545.

Au, K.F., Sebastiano, V., Afshar, P.T., Durruthy, J.D., Lee, L., Williams, B.A., van Bakel, H., Schadt, E.E., Reijo-Pera, R.A., Underwood, J.G., et al. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. U. S. A. *110*, E4821–E4830.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Clementi, L., Ren, J., Li, W., and Noble, W. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Bakkour, N., Lin, Y.-L., Maire, S., Ayadi, L., Mahuteau-Betzer, F., Nguyen, C.H., Mettling, C., Portales, P., Grierson, D., Chabot, B., et al. (2007). Small-Molecule Inhibition of HIV pre-mRNA Splicing as a Novel Antiretroviral Therapy to Overcome Drug Resistance. PLoS Pathog. *3*.

Balwierz, P.J., Carninci, P., Daub, C.O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., and van Nimwegen, E. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. *10*, R79.

Balwierz, P.J., Pachkov, M., Arnold, P., Gruber, A.J., Zavolan, M., and van Nimwegen, E. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res. *24*, 869–884.

Barash, Y., and Vaquero-Garcia, J. (2014). Splicing Code Modeling. In Systems Biology of RNA Binding Proteins, G.W. Yeo, ed. (New York, NY: Springer New York), pp. 451–466.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science *338*, 1587–1593.

Bauer, D.C., Buske, F.A., and Bailey, T.L. (2010). Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. BMC Bioinformatics *11*, 366.

Becattini, S., Latorre, D., Mele, F., Foglierini, M., De Gregorio, C., Cassotta, A., Fernandez, B., Kelderman, S., Schumacher, T.N., Corti, D., et al. (2015). T cell immunity. Functional heterogeneity of human memory $CD4^+$ T cell clones primed by pathogens or vaccines. Science *347*, 400–406.

Becskei, A., Séraphin, B., and Serrano, L. (2001). Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. EMBO J. *20*, 2528–2535.

Bell, N.M., and Lever, A.M.L. (2013). HIV Gag polyprotein: processing and early viral particle assembly. Trends Microbiol. *21*, 136–144.

Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu. Rev. Biochem. *72*, 291–336.

Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends Biochem. Sci. *42*, 407–408.

Bonilla, F.A., and Oettgen, H.C. (2010). Adaptive immunity. J. Allergy Clin. Immunol. *125*, S33-40.

Braunschweig, U., Gueroussov, S., Plocik, A.M., Graveley, B.R., and Blencowe, B.J. (2013). Dynamic Integration of Splicing within Gene Regulatory Pathways. Cell *152*, 1252–1269.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Brenchley, J.M., Douek, D.C., Ambrozak, D.R., Chatterji, M., Betts, M.R., Davis, L.S., and Koup, R.A. (2002). Expansion of activated human naïve T-cells precedes effector function. Clin. Exp. Immunol. *130*, 431–440.

Brierley, I., and Dos Ramos, F.J. (2006). Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. Virus Res. *119*, 29–42.

Brugiolo, M., Herzel, L., and Neugebauer, K.M. (2013). Counting on co-transcriptional splicing. F1000prime Rep. *5*, 9.

Budach, S., and Marsico, A. (2018). pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinforma. Oxf. Engl. *34*, 3035–3037.

Buljan, M., Chalancon, G., Dunker, A.K., Bateman, A., Balaji, S., Fuxreiter, M., and Babu, M.M. (2013). Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. Curr. Opin. Struct. Biol. *23*, 443–450.

Burge, C., Graveley, B., Yeo, G.W., Alexis, M.S., Freese, P., and Dominguez, D. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell *70*, 854-867.e9.

Burset, M., Seledtsov, I.A., and Solovyev, V.V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. *28*, 4364–4375.

Busch, A., and Hertel, K.J. (2015). Splicing predictions reliably classify different types of alternative splicing. RNA *21*, 813–823.

Busslinger, M., Moschonas, N., and Flavell, R.A. (1981). Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. Cell *27*, 289–298.

Butte, M.J., Lee, S.J., Jesneck, J., Keir, M.E., Haining, W.N., and Sharpe, A.H. (2012). CD28 costimulation regulates genome-wide effects on alternative splicing. PloS One *7*, e40032.

Carazo, F., Romero, J.P., and Rubio, A. (2018). Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. Brief. Bioinform. *20*, 1358–1375.

Cen, S., Peng, Z.-G., Li, X.-Y., Li, Z.-R., Ma, J., Wang, Y.-M., Fan, B., You, X.-F., Wang, Y.-P., Liu, F., et al. (2010). Small molecular compounds inhibit HIV-1 replication through specifically stabilizing APOBEC3G. J. Biol. Chem. *285*, 16546–16552.

Chawla, G., Lin, C.-H., Han, A., Shiue, L., Ares, M., and Black, D.L. (2009). Sam68 Regulates a Set of Alternatively Spliced Exons during Neurogenesis. Mol. Cell. Biol. *29*, 201–213.

Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999). Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. Genes Dev. *13*, 593–606.

Chen, L., Kostadima, M., Martens, J.H.A., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033.

Chou, M.Y., Rooke, N., Turck, C.W., and Black, D.L. (1999). hnRNP H is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. Mol. Cell. Biol. *19*, 69–77.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F. (2005). A role for LEDGF/p75 in targeting HIV DNA integration. Nat. Med. *11*, 1287–1289.

Cole, B.S., Tapescu, I., Allon, S.J., Mallory, M.J., Qiu, J., Lake, R.J., Fan, H.-Y., Fu, X.-D., and Lynch, K.W. (2015). Global analysis of physical and functional RNA targets of hnRNP L reveals distinct sequence and epigenetic features of repressed and enhanced exons. RNA *21*, 2053–2066.

Compton, A.A., Bruel, T., Porrot, F., Mallet, A., Sachse, M., Euvrard, M., Liang, C., Casartelli, N., and Schwartz, O. (2014). IFITM proteins incorporated into HIV-1 virions impair viral fusion and spread. Cell Host Microbe *16*, 736–747.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Courtney, A.H., Lo, W.-L., and Weiss, A. (2018). TCR Signaling: Mechanisms of Initiation and Propagation. Trends Biochem. Sci. *43*, 108–123.

Courtney, D.G., Tsai, K., Bogerd, H.P., Kennedy, E.M., Law, B.A., Emery, A., Swanstrom, R., Holley, C.L., and Cullen, B.R. (2019). Epitranscriptomic Addition of m5C to HIV-1 Transcripts Regulates Viral Gene Expression. Cell Host Microbe *26*, 217-227.e6.

Craigie, R., and Bushman, F.D. (2012). HIV DNA Integration. Cold Spring Harb. Perspect. Med. *2*.

Damgaard, C.K., Tange, T.Ø., and Kjems, J. (2002). hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. RNA *8*, 1401–1415.

Danan-Gotthold, M., Golan-Gerstl, R., Eisenberg, E., Meir, K., Karni, R., and Levanon, E.Y. (2015). Identification of recurrent regulated alternative splicing events across human solid tumors. Nucleic Acids Res. *43*, 5130–5144.

Das, D., Clark, T.A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J.E., et al. (2007). A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic Acids Res. *35*, 4845–4857.

David, C.J., and Manley, J.L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. *24*, 2343–2364.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Ding, L., Rath, E., and Bai, Y. (2017). Comparison of Alternative Splicing Junction Detection Tools Using RNA-Seq Data. Curr. Genomics *18*, 268–277.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Djuretic, I.M., Levanon, D., Negreanu, V., Groner, Y., Rao, A., and Ansel, K.M. (2007). Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells. Nat. Immunol. *8*, 145–153.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doyle, T., Goujon, C., and Malim, M.H. (2015). HIV-1 and interferons: who's interfering with whom? Nat. Rev. Microbiol. *13*, 403–413.

DuPage, M., and Bluestone, J.A. (2016). Harnessing the plasticity of CD4+ T cells to treat immune-mediated disease. Nat. Rev. Immunol. *16*, 149–163.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Dvinge, H. (2018). Regulation of alternative mRNA splicing: old players and new perspectives. FEBS Lett. *592*, 2987–3006.

Erkelenz, S., Mueller, W.F., Evans, M.S., Busch, A., Schöneweis, K., Hertel, K.J., and Schaal, H. (2013). Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. RNA *19*, 96–102.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

Ezkurdia, I., Rodriguez, J.M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A., and Tress, M.L. (2015). Most highly expressed protein-coding genes have a single dominant isoform. J. Proteome Res. *14*, 1880–1887.

Feng, H., Bao, S., Rahman, M.A., Weyn-Vanhentenryck, S.M., Khan, A., Wong, J., Shah, A., Flynn, E.D., Krainer, A.R., and Zhang, C. (2019). Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. Mol. Cell *74*, 1189-1204.e6.

Feng, Y., Broder, C.C., Kennedy, P.E., and Berger, E.A. (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science *272*, 872–877.

Feng, Y.-Y., Ramu, A., Cotto, K.C., Skidmore, Z.L., Kunisaki, J., Conrad, D.F., Lin, Y., Chapman, W.C., Uppaluri, R., Govindan, R., et al. (2018). RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. BioRxiv 436634.

Ficarelli, M., Wilson, H., Pedro Galão, R., Mazzon, M., Antzin-Anduetza, I., Marsh, M., Neil, S.J., and Swanson, C.M. (2019). KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. ELife *8*, e46767.

Ficarelli, M., Antzin-Anduetza, I., Hugh-White, R., Firth, A.E., Sertkaya, H., Wilson, H., Neil, S.J.D., Schulz, R., and Swanson, C.M. (2020). CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. J. Virol. *94*.

Fros, J.J., Dietrich, I., Alshaikhahmed, K., Passchier, T.C., Evans, D.J., and Simmonds, P. (2017). CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. ELife *6*, e29112.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Fu, K., Sun, X., Zheng, W., Wier, E.M., Hodgson, A., Tran, D.Q., Richard, S., and Wan, F. (2013). Sam68 modulates the promoter specificity of NF-κB and mediates expression of CD25 in activated T cells. Nat. Commun. *4*, 1909.

Fusaki, N., Iwamatsu, A., Iwashima, M., and Fujisawa, J. i (1997). Interaction between Sam68 and Src family tyrosine kinases, Fyn and Lck, in T cell receptor signaling. J. Biol. Chem. *272*, 6214–6219.

Galganski, L., Urbanek, M.O., and Krzyzosiak, W.J. (2017). Nuclear speckles: molecular organization, biological function and role in disease. Nucleic Acids Res. *45*, 10350–10368.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257–2267.

Gaunt, E., Wise, H.M., Zhang, H., Lee, L.N., Atkinson, N.J., Nicol, M.Q., Highton, A.J., Klenerman, P., Beard, P.M., Dutia, B.M., et al. (2016). Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. ELife *5*, e12735.

Gehring, N.H., Lamprinaki, S., Kulozik, A.E., and Hentze, M.W. (2009). Disassembly of exon junction complexes by PYM. Cell *137*, 536–548.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Geuens, T., Bouhy, D., and Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. Hum. Genet. *135*, 851.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput. Biol. *10*, e1003711.

Giudice, G., Sánchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. Database J. Biol. Databases Curation *2016*.

Glória, V.G. da, Araújo, M.M. de, Santos, A.M., Leal, R., Almeida, S.F. de, Carmo, A.M., and Moreira, A. (2014). T Cell Activation Regulates CD6 Alternative Splicing by Transcription Dynamics and SRSF1. J. Immunol. *193*, 391–399.

Goodwin, R.G., Friend, D., Ziegler, S.F., Jerzy, R., Falk, B.A., Gimpel, S., Cosman, D., Dower, S.K., March, C.J., and Namen, A.E. (1990). Cloning of the human and murine interleukin-7 receptors: demonstration of a soluble form and homology to a new receptor superfamily. Cell *60*, 941–951.

Goujon, C., Moncorge, O., Bauby, H., Doyle, T., Ward, C., Schaller, T., Hue, S., Barclay, W., Schulz, R., and Malim, M. (2013). Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. Nature *502*, 559–562.

Gray, J.T., Lee, J.-S., and Mulligan, R.C. (2005). Packaging cells comprising codon-optimized gagpol sequences and lacking lentiviral accessory proteins.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Hahne, F., and Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In Statistical Genomics, E. Mathé, and S. Davis, eds. (Springer New York), pp. 335–351.

Han, K., Yeo, G., An, P., Burge, C.B., and Grabowski, P.J. (2005). A Combinatorial Code for Splicing Silencing: UAGG and GGGG Motifs. PLOS Biol. *3*, e158.

Harvey, S.E., and Cheng, C. (2016). Methods for Characterization of Alternative RNA Splicing. Methods Mol. Biol. Clifton NJ *1402*, 229–241.

Havlioglu, N., Wang, J., Fushimi, K., Vibranovski, M.D., Kan, Z., Gish, W., Fedorov, A., Long, M., and Wu, J.Y. (2007). An intronic signal for alternative splicing in the human genome. PloS One *2*, e1246.

Hawse, W.F., Boggess, W.C., and Morel, P.A. (2017). TCR Signal Strength Regulates Akt Substrate Specificity To Induce Alternate Murine Th and T Regulatory Cell Differentiation Programs. J. Immunol. ji1700369.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies11Edited by G. von Heijne. J. Mol. Biol. *281*, 827–842.

Henriksson, J., Chen, X., Gomes, T., Ullah, U., Meyer, K.B., Miragaia, R., Duddy, G., Pramanik, J., Yusa, K., Lahesmaa, R., et al. (2019). Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. Cell *176*, 882-896.e18.

Hermiston, M.L., Xu, Z., and Weiss, A. (2003). CD45: A Critical Regulator of Signaling Thresholds in Immune Cells. Annu. Rev. Immunol. *21*, 107–137.

Hertel, K.J. (2008). Combinatorial Control of Exon Recognition. J. Biol. Chem. *283*, 1211–1215.

Heyd, F., and Lynch, K.W. (2010). Phosphorylation-Dependent Regulation of PSF by GSK3 controls CD45 Alternative Splicing. Mol. Cell *40*, 126–137.

Hooper, J.E. (2014). A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Hum. Genomics *8*, 3.

House, A.E., and Lynch, K.W. (2006). An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. Nat. Struct. Mol. Biol. *13*, 937–944.

Hsieh, C.-S., Lee, H.-M., and Lio, C.-W.J. (2012). Selection of regulatory T cells in the thymus. Nat. Rev. Immunol. *12*, 157–167.

Hu, W.-S., and Hughes, S.H. (2012). HIV-1 Reverse Transcription. Cold Spring Harb. Perspect. Med. *2*.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., et al. (2012). Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins. Cell Rep. *1*, 167–178.

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. *24*, 1988–1998.

Hunt, S.E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I.M., Trevanion, S.J., Flicek, P., et al. (2018). Ensembl variation resources. Database *2018*.

Huot, M.-É., Vogel, G., Zabarauskas, A., Ngo, C.T.-A., Coulombe-Huntington, J., Majewski, J., and Richard, S. (2012). The Sam68 STAR RNA-binding protein regulates mTOR alternative splicing during adipogenesis. Mol. Cell *46*, 187–199.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein–RNA interactions at nucleotide resolution. Methods San Diego Calif *65*, 274–287.

Hwang, J., and Kim, Y.K. (2013). When a ribosome encounters a premature termination codon. BMB Rep. *46*, 9–16.

Iancu, O.D., Colville, A., Oberbeck, D., Darakjian, P., McWeeney, S.K., and Hitzemann, R. (2015). Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations. Front. Genet. *6*.

Inada, T. (2017). The Ribosome as a Platform for mRNA and Nascent Polypeptide Quality Control. Trends Biochem. Sci. *42*, 5–15.

Ip, J.Y., Tong, A., Pan, Q., Topp, J.D., Blencowe, B.J., and Lynch, K.W. (2007). Global analysis of alternative splicing during T-cell activation. RNA *13*, 563–572.

Irimia, M., and Roy, S.W. (2014). Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harb. Perspect. Biol. *6*.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell *159*, 1511–1523.

Isken, O., Kim, Y.K., Hosoda, N., Mayeur, G.L., Hershey, J.W.B., and Maquat, L.E. (2008). Upf1 Phosphorylation Triggers Translational Repression during Nonsense-Mediated mRNA Decay. Cell *133*, 314–327.

Itakura, A.K., Futia, R.A., and Jarosz, D.F. (2018). It Pays To Be In Phase. Biochemistry *57*, 2520–2529.

Izquierdo, J.M., and Valcárcel, J. (2007). Fas-activated Serine/Threonine Kinase (FAST K) Synergizes with TIA-1/TIAR Proteins to Regulate Fas Alternative Splicing. J. Biol. Chem. *282*, 1539–1543.

Jangi, M., and Sharp, P.A. (2014). Building Robust Transcriptomes with Master Splicing Factors. Cell *159*, 487–498.

Jangi, M., Boutz, P.L., Paul, P., and Sharp, P.A. (2014). Rbfox2 controls autoregulation in RNA-binding protein networks. Genes Dev. *28*, 637–651.

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. Mol. Cells *40*, 1.

Jha, A., Gazzara, M.R., and Barash, Y. (2017). Integrative deep models for alternative splicing. Bioinformatics *33*, i274–i282.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Jolma, A., Zhang, J., Mondragón, E., Kivioja, T., Yin, Y., Zhu, F., Morris, Q., Hughes, T.R., Maher, L.J., and Taipale, J. (2019). Binding specificities of human RNA binding proteins towards structured and linear RNA sequences. BioRxiv 317909.

Kahles, A., Ong, C.S., Zhong, Y., and Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics *32*, 1840–1847.

Kalekar, L.A., Schmiel, S.E., Nandiwada, S.L., Lam, W.Y., Barsness, L.O., Zhang, N., Stritesky, G.L., Malhotra, D., Pauken, K.E., Linehan, J.L., et al. (2016). CD4+ T cell anergy prevents autoimmunity and generates regulatory T cell precursors. Nat. Immunol. *17*, 304–314.

Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat. Rev. Genet. *12*, 715–729.

Karlin, S., Doerfler, W., and Cardon, L.R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. *68*, 2889–2897.

Karn, J., and Stoltzfus, C.M. (2012). Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. Cold Spring Harb. Perspect. Med. *2*, a006916.

Ke, S., and Chasin, L.A. (2010). Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. Genome Biol. *11*, R84.

Keene, J.D., Komisarow, J.M., and Friedersdorf, M.B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat. Protoc. *1*, 302–307.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene *514*, 1–30.

Khaldoyanidi, S., Achtnich, M., Hehlmann, R., and Zöller, M. (1996). Expression of CD44 variant isoforms in peripheral blood leukocytes in malignant lymphoma and leukemia: Inverse correlation between expression and tumor progression. Leuk. Res. *20*, 839–851.

Kim, C.H., Campbell, D.J., and Butcher, E.C. (2001). Nonpolarized memory T cells. Trends Immunol. *22*, 527–530.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. *35*, 125–131.

Kim, H.P., Imbert, J., and Leonard, W.J. (2006). Both integrated and differential regulation of components of the IL-2/IL-2 receptor system. Cytokine Growth Factor Rev. *17*, 349–366.

Klatzmann, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., Gluckman, J.C., and Montagnier, L. (1984). T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. Nature *312*, 767–768.

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12.

Kondrack, R.M., Harbertson, J., Tan, J.T., McBreen, M.E., Surh, C.D., and Bradley, L.M. (2003). Interleukin 7 regulates the survival and generation of memory CD4 cells. J. Exp. Med. *198*, 1797–1806.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat. Rev. Mol. Cell Biol. *14*, 153–165.

Krawczak, M., Reiss, J., and Cooper, D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet. *90*, 41–54.

Kreslavsky, T., Gleimer, M., Garbe, A.I., and von Boehmer, H. (2010). αβ versus γδ fate choice: counting the T-cell lineages at the branch point. Immunol. Rev. *238*, 169–181.

Kunkel, G.T., and Wang, X. (2011). Sam68 guest STARs in TNF-α signaling. Mol. Cell *43*, 157–158.

Kypr, J., Mrázek, J., and Reich, J. (1989). Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 12. Biochim. Biophys. Acta BBA - Gene Struct. Expr. *1009*, 280–282.

La Porta, J., Matus-Nicodemos, R., Valentín-Acevedo, A., and Covey, L.R. (2016). The RNA-Binding Protein, Polypyrimidine Tract-Binding Protein 1 (PTBP1) Is a Key Regulator of CD4 T Cell Activation. PLoS ONE *11*.

La Rosa, P., Bielli, P., Compagnucci, C., Cesari, E., Volpe, E., Farioli Vecchioli, S., and Sette, C. (2016). Sam68 promotes self-renewal and glycolytic metabolism in mouse neural progenitor cells by modulating Aldh1a3 pre-mRNA 3'-end processing. ELife *5*, e20750.

LaMere, S.A., Thompson, R.C., Komori, H.K., Mark, A., and Salomon, D.R. (2016). Promoter H3K4 methylation dynamically reinforces activation-induced pathways in human CD4 T cells. Genes Immun. *17*, 283–297.

LaMere, S.A., Thompson, R.C., Meng, X., Komori, H.K., Mark, A., and Salomon, D.R. (2017). H3K27 methylation dynamics during CD4 T cell activation: regulation of JAK/STAT and IL12RB2 expression by JMJD3. J. Immunol. Baltim. Md 1950 *199*, 3158–3175.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719–720.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926–929.

Lauer, U.M., Staehler, P., Lambrecht, R.M., Oberdorfer, F., Spiegel, M., Wybranietz, W.A., Gross, C.D., and Gregor, M. (2000). A prototype transduction tag system (delta LNGFR/NGF) for noninvasive clinical gene therapy monitoring. Cancer Gene Ther. *7*, 430–437.

Lavender, C.A., Gorelick, R.J., and Weeks, K.M. (2015). Structure-Based Alignment and Consensus Secondary Structures for Three HIV-Related RNA Genomes. PLoS Comput. Biol. *11*, e1004230.

Le Grice, S.F.J. (2012). Human immunodeficiency virus reverse transcriptase: 25 years of research, drug discovery, and promise. J. Biol. Chem. *287*, 40850–40857.

LeBien, T.W., and Tedder, T.F. (2008). B lymphocytes: how they develop and function. Blood *112*, 1570–1580.

LeBlanc, J., Weil, J., and Beemon, K. (2013). Posttranscriptional regulation of retroviral gene expression: primary RNA transcripts play three roles as pre-mRNA, mRNA, and genomic RNA. WIREs RNA *4*, 567–580.

Lee, C., and Roy, M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol. *5*, 231.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21.

Lemaire, R., Winne, A., Sarkissian, M., and Lafyatis, R. (1999). SF2 and SRp55 regulation of CD45 exon 4 skipping during T cell activation. Eur. J. Immunol. *29*, 823–837.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. *98*, 11193–11198.

Liu, T., and Lin, K. (2015). The distribution pattern of genetic variation in the transcript isoforms of the alternatively spliced protein-coding genes in the human genome. Mol. Biosyst. *11*, 1378–1388.

Liu, C., Cheng, J., and Mountz, J.D. (1995). Differential expression of human Fas mRNA species upon peripheral blood mononuclear cell activation. Biochem. J. *310*, 957–963.

Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R., and Wickramasinghe, V.O. (2017). Impact of Alternative Splicing on the Human Proteome. Cell Rep. *20*, 1229–1241.

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., et al. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat. Struct. Mol. Biol. *17*, 1114–1123.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. *579*, 1900–1903.

Louadi, Z., Oubounyt, M., Tayara, H., and Chong, K.T. (2019). Deep Splicing Code: Classifying Alternative Splicing Events Using Deep Learning. Genes *10*.

Lu, J., Pan, Q., Rong, L., He, W., Liu, S.-L., and Liang, C. (2011). The IFITM proteins inhibit HIV-1 infection. J. Virol. *85*, 2126–2137.

Luckheeram, R.V., Zhou, R., Verma, A.D., and Xia, B. (2012). $CD4^+$T cells: differentiation and functions. Clin. Dev. Immunol. *2012*, 925135.

Madsen, J.G.S., Rauch, A., Hauwaert, E.L.V., Schmidt, S.F., Winnefeld, M., and Mandrup, S. (2018). Integrated analysis of motif activity and gene expression changes of transcription factors. Genome Res. *28*, 243–255.

Maggi, J., Schafer, C., Ubilla-Olguín, G., Catalán, D., Schinnerling, K., and Aguillón, J.C. (2015). Therapeutic potential of hyporesponsive CD4+ T cells in autoimmunity. Immunol. Toler. 488.

Mahiet, C., and Swanson, C.M. (2016). Control of HIV-1 gene expression by SR proteins. Biochem. Soc. Trans. *44*, 1417–1425.

Malek, T.R., and Castro, I. (2010). Interleukin-2 Receptor Signaling: At the Interface between Tolerance and Immunity. Immunity *33*, 153–165.

Mamik, M.K., and Ghorpade, A. (2014). Chemokine CXCL8 Promotes HIV-1 Replication in Human Monocyte-Derived Macrophages and Primary Microglia via Nuclear Factor-κB Pathway. PLoS ONE *9*.

Mandel, J. (1982). Use of the Singular Value Decomposition in Regression Analysis. Am. Stat. *36*, 15–24.

Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C.M., and Tartaglia, G.G. (2016). Advances in the characterization of RNA-binding proteins. Wiley Interdiscip. Rev. RNA *7*, 793–810.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martinez, N.M., and Lynch, K.W. (2013). Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. Immunol. Rev. *253*, 216–236.

Martinez, N.M., Pan, Q., Cole, B.S., Yarosh, C.A., Babcock, G.A., Heyd, F., Zhu, W., Ajith, S., Blencowe, B.J., and Lynch, K.W. (2012). Alternative splicing networks regulated by signaling in human T cells. RNA *18*, 1029–1040.

Martinez, N.M., Agosto, L., Qiu, J., Mallory, M.J., Gazzara, M.R., Barash, Y., Fu, X.-D., and Lynch, K.W. (2015). Widespread JNK-dependent alternative splicing induces a positive feedback loop through CELF2-mediated regulation of MKK7 during T-cell activation. Genes Dev. *29*, 2054–2066.

Martínez-Méndez, D., Villarreal, C., Mendoza, L., and Huerta, L. (2020). An Integrative Network Modeling Approach to T CD4 Cell Activation. Front. Physiol. *11*.

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N., and Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. Genes Dev. *19*, 1512–1517.

Maticzka, D., Lange, S.J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. *15*, R17.

Matter, N., Herrlich, P., and König, H. (2002). Signal-dependent regulation of splicing via phosphorylation of Sam68. Nature *420*, 691–695.

Mauger, D.M., Lin, C., and Garcia-Blanco, M.A. (2008). hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. Mol. Cell. Biol. *28*, 5403–5419.

Mayrose, I., Stern, A., Burdelova, E.O., Sabo, Y., Laham-Karam, N., Zamostiano, R., Bacharach, E., and Pupko, T. (2013). Synonymous site conservation in the HIV-1 genome. BMC Evol. Biol. *13*, 164.

McNeill, L., Cassady, R.L., Sarkardei, S., Cooper, J.C., Morgan, G., and Alexander, D.R. (2004). CD45 isoforms in T cell signalling and development. Immunol. Lett. *92*, 125–134.

Meininger, I., Griesbach, R.A., Hu, D., Gehring, T., and Krappmann, D. (2016). Alternative splicing of MALT1 controls signalling and activation of CD4+ T cells. Nat. Commun. *7*.

Melamud, E., and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–4886.

Melton, A.A., Jackson, J., Wang, J., and Lynch, K.W. (2007). Combinatorial Control of Signal-Induced Exon Repression by hnRNP L and PSF. Mol. Cell. Biol. *27*, 6972–6984.

Mercier, E., and Gottardo, R. (2018). MotIV: Motif Identification and Validation. R package version 1.39.0.

Merkin, J., Russell, C., Chen, P., and Burge, C. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Science *338*, 1593–1599.

Mi, Z., Ding, J., Zhang, Q., Zhao, J., Ma, L., Yu, H., Liu, Z., Shan, G., Li, X., Zhou, J., et al. (2015). A small molecule compound IMB-LA inhibits HIV-1 infection by preventing viral Vpu from antagonizing the host restriction factor BST-2. Sci. Rep. *5*, 18499.

Miyakawa, K., Matsunaga, S., Kanou, K., Matsuzawa, A., Morishita, R., Kudoh, A., Shindo, K., Yokoyama, M., Sato, H., Kimura, H., et al. (2015). ASK1 restores the antiviral activity of APOBEC3G by disrupting HIV-1 Vif-mediated counteraction. Nat. Commun. *6*, 6945.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet. *30*, 13–19.

Moldón, A., and Query, C. (2010). Crossing the Exon. Mol. Cell *38*, 159–161.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Motta-Mena, L.B., Heyd, F., and Lynch, K.W. (2010). Context-dependent regulatory mechanism of the splicing factor hnRNP L. Mol. Cell *37*, 223.

Moulton, V.R., and Tsokos, G.C. (2010). Alternative splicing factor/splicing factor 2 regulates the expression of the zeta subunit of the human T cell receptor-associated CD3 complex. J. Biol. Chem. *285*, 12490–12496.

Mount, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res. *10*, 459–472.

Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem. Sci. *23*, 198–199.

Neil, S.J.D. (2013). The antiviral activities of tetherin. Curr. Top. Microbiol. Immunol. *371*, 67–104.

Neil, S.J.D., Zang, T., and Bieniasz, P.D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature *451*, 425–430.

Nevozhay, D., Adams, R.M., Murphy, K.F., Josic, K., and Balázsi, G. (2009). Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. Proc. Natl. Acad. Sci. U. S. A. *106*, 5123–5128.

Ngandu, N.K., Scheffler, K., Moore, P., Woodman, Z., Martin, D., and Seoighe, C. (2008). Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol. J. *5*, 160.

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., et al. (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res. *46*, W209–W214.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21*, 708–718.

Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., Zhou, Z., Dai, Y., Yang, Y., Liu, P., et al. (2016). Global intron retention mediated gene regulation during CD4+ T cell activation. Nucleic Acids Res. gkw591.

Nilsen, T., and Graveley, B. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526–540.

Nostrand, E.L.V., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2018). A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. BioRxiv 179648.

Oberdoerffer, S., Moita, L.F., Neems, D., Freitas, R.P., Hacohen, N., and Rao, A. (2008). Regulation of CD45 Alternative Splicing by Heterogeneous Ribonucleoprotein, hnRNPLL. Science *321*, 686–691.

Okoye, A.A., and Picker, L.J. (2013). CD4+ T cell depletion in HIV infection: mechanisms of immunological failure. Immunol. Rev. *254*, 54–64.

Okunola, H.L., and Krainer, A.R. (2009). Cooperative-Binding and Splicing-Repressive Properties of hnRNP A1. Mol. Cell. Biol. *29*, 5620–5631.

Orenstein, Y., Wang, Y., and Berger, B. (2016). RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data. Bioinformatics *32*, i351–i359.

O'Shea, J.J., and Paul, W.E. (2010). Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. Science *327*, 1098–1102.

Ou, J., Wolfe, S.A., Brodsky, M.H., and Zhu, L.J. (2018). motifStack for the analysis of transcription factor binding site evolution. Nat. Methods *15*, 8–9.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. *20*, 153–158.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M., Jr, et al. (2013). Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. Mol. Cell *50*, 223.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. *102*, 11–26.

Paronetto, M.P., Venables, J.P., Elliott, D.J., Geremia, R., Rossi, P., and Sette, C. (2003). Tr-kit promotes the formation of a multimolecular complex composed by Fyn, PLCgamma1 and Sam68. Oncogene *22*, 8707–8715.

Paronetto, M.P., Achsel, T., Massiello, A., Chalfant, C.E., and Sette, C. (2007). The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. J. Cell Biol. *176*, 929–939.

Paronetto, M.P., Messina, V., Bianchi, E., Barchi, M., Vogel, G., Moretti, C., Palombi, F., Stefanini, M., Geremia, R., Richard, S., et al. (2009). Sam68 regulates translation of target mRNAs in male germ cells, necessary for mouse spermatogenesis. J. Cell Biol. *185*, 235–249.

Paronetto, M.P., Messina, V., Barchi, M., Geremia, R., Richard, S., and Sette, C. (2011). Sam68 marks the transcriptionally active stages of spermatogenesis and modulates alternative splicing in male germ cells. Nucleic Acids Res. *39*, 4961–4974.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Paz, I., Kosti, I., Ares, M., Cline, M., and Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. *42*, W361–W367.

Pedrotti, S., Bielli, P., Paronetto, M.P., Ciccosanti, F., Fimia, G.M., Stamm, S., Manley, J.L., and Sette, C. (2010). The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy. EMBO J. *29*, 1235–1247.

Pettit, S.C., Moody, M.D., Wehbie, R.S., Kaplan, A.H., Nantermet, P.V., Klein, C.A., and Swanstrom, R. (1994). The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. J. Virol. *68*, 8017–8027.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690.

Plaschka, C., Newman, A.J., and Nagai, K. (2019). Structural Basis of Nuclear pre-mRNA Splicing: Lessons from Yeast. Cold Spring Harb. Perspect. Biol. *11*, a032391.

Plocik, A.M., and Guthrie, C. (2012). Diverse forms of RPS9 splicing are part of an evolving autoregulatory circuit. PLoS Genet. *8*, e1002620.

Pollard, V.W., and Malim, M.H. (1998). The HIV-1 Rev protein. Annu. Rev. Microbiol. *52*, 491–532.

Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R., and Bickmore, W.A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. PLoS Genet. *8*, e1002717.

Pyle, A.M. (2016). Group II Intron Self-Splicing. Annu. Rev. Biophys. *45*, 183–205.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinforma. *47*, 11.12.1-11.12.34.

R Core Team (2019). R Core Team (2019). R: A language and environment for statistical computing.

Raj, B., O'Hanlon, D., Vessey, J.P., Pan, Q., Ray, D., Buckley, N.J., Miller, F.D., and Blencowe, B.J. (2011). Cross-Regulation between an Alternative Splicing Activator and a Transcription Repressor Controls Neurogenesis. Mol. Cell *43*, 843–850.

Ramakrishnan, P., and Baltimore, D. (2011). Sam68 Is Required for Both NF-κB Activation and Apoptosis Signaling by the TNF Receptor. Mol. Cell *43*, 167–179.

Randau, L., and Söll, D. (2008). Transfer RNA genes in pieces. EMBO Rep. *9*, 623–628.

Rathmell, J.C., and Thompson, C.B. (2002). Pathways of apoptosis in lymphocyte development, homeostasis, and disease. Cell *109 Suppl*, S97-107.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. *27*, 667–670.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. *35*, W193–W200.

Reimand, J., Kolde, R., and Arak, T. (2018). gProfileR: Interface to the "g:Profiler" Toolkit. R package version 0.6.7.

Reinhold-Hurek, B., and Shub, D.A. (1992). Self-splicing introns in tRNA genes of widely divergent bacteria. Nature *357*, 173–176.

Rice, A.P. (2017). The HIV-1 Tat protein: mechanism of action and target for HIV-1 cure strategies. Curr. Pharm. Des. *23*, 4098–4102.

Riley, T.R., Lazarovici, A., Mann, R.S., and Bussemaker, H.J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. ELife *4*.

Rima, B.K., and McFerran, N.V. (1997). Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J. Gen. Virol. *78 (Pt 11)*, 2859–2870.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. gkv007.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Rock, K.L., Reits, E., and Neefjes, J. (2016). Present Yourself! By MHC Class I and MHC Class II Molecules. Trends Immunol. *37*, 724–737.

Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., et al. (2012). Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci. Rep. *2*, 603.

Rollins, C., Levengood, J.D., Rife, B.D., Salemi, M., and Tolbert, B.S. (2014). Thermodynamic and Phylogenetic Insights into hnRNP A1 Recognition of the HIV-1 Exon Splicing Silencer 3 Element. Biochemistry *53*, 2172–2184.

Rosa, A., Chande, A., Ziglio, S., De Sanctis, V., Bertorelli, R., Goh, S.L., McCauley, S.M., Nowosielska, A., Antonarakis, S.E., Luban, J., et al. (2015). HIV-1 Nef promotes infection by excluding SERINC5 from virion incorporation. Nature *526*, 212–217.

Roth, D.B. (2014). V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol. Spectr. *2*.

Rothrock, C.R., House, A.E., and Lynch, K.W. (2005). HnRNP L represses exon splicing via a regulated exonic splicing silencer. EMBO J. *24*, 2792–2802.

Ryder, S.P., Recht, M.I., and Williamson, J.R. (2008). Quantitative analysis of protein-RNA interactions by gel mobility shift. Methods Mol. Biol. Clifton NJ *488*, 99–115.

Sakharkar, M.K., Chow, V.T.K., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol. *4*, 387–393.

Salmena, L., Lemmers, B., Hakem, A., Matysiak-Zablocki, E., Murakami, K., Au, P.Y.B., Berry, D.M., Tamblyn, L., Shehabeldin, A., Migon, E., et al. (2003). Essential role for caspase 8 in T-cell homeostasis and T-cell-mediated immunity. Genes Dev. *17*, 883–895.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. *28*, 4320–4330.

Sanford, J.R., Gray, N.K., Beckmann, K., and Cáceres, J.F. (2004). A novel role for shuttling SR proteins in mRNA translation. Genes Dev. *18*, 755–768.

Saravia, J., Chapman, N.M., and Chi, H. (2019). Helper T cell differentiation. Cell. Mol. Immunol. *16*, 634–643.

Sarzotti-Kelsoe, M., Bailer, R.T., Turk, E., Lin, C., Bilska, M., Greene, K.M., Gao, H., Todd, C.A., Ozaki, D.A., Seaman, M.S., et al. (2014). Optimization and Validation of the TZM-bl Assay for Standardized Assessments of Neutralizing Antibodies Against HIV-1. J. Immunol. Methods *0*, 131–146.

Sasse, A., Laverty, K.U., Hughes, T.R., and Morris, Q.D. (2018). Motif models for RNA-binding proteins. Curr. Opin. Struct. Biol. *53*, 115–123.

Schmidl, C., Delacher, M., Huehn, J., and Feuerer, M. (2018). Epigenetic mechanisms regulating T-cell responses. J. Allergy Clin. Immunol. *142*, 728–743.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. Cell *101*, 671–684.

Schoenberg, D.R., and Maquat, L.E. (2012). Regulation of cytoplasmic mRNA decay. Nat. Rev. Genet. *13*, 246–259.

Schwarz, B.A., and Bhandoola, A. (2006). Trafficking from the bone marrow to the thymus: a prerequisite for thymopoiesis. Immunol. Rev. *209*, 47–57.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. *26*, 732–744.

Sertznig, H., Hillebrand, F., Erkelenz, S., Schaal, H., and Widera, M. (2018). Behind the scenes of HIV-1 replication: Alternative splicing as the dependency factor on the quiet. Virology *516*, 176–188.

Shankarling, G., Cole, B.S., Mallory, M.J., and Lynch, K.W. (2014). Transcriptome-Wide RNA Interaction Profiling Reveals Physical and Functional Targets of hnRNP L in Human T Cells. Mol. Cell. Biol. *34*, 71–83.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014.

Sharp, P.A. (1994). Split genes and RNA splicing. Cell *77*, 805–815.

Sharp, P.M., and Hahn, B.H. (2011). Origins of HIV and the AIDS Pandemic. Cold Spring Harb. Perspect. Med. *1*.

Shen, H., and Green, M.R. (2006). RS domains contact splicing signalsand promote splicing by a commonmechanism in yeast through humans. Genes Dev. *20*, 1755–1765.

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593-5601.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat. Rev. Mol. Cell Biol. *18*, 655–670.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.

Si, Z., Rauch, D., and Stoltzfus, C.M. (1998). The Exon Splicing Silencer in Human Immunodeficiency Virus Type 1 Tat Exon 3 Is Bipartite and Acts Early in Spliceosome Assembly. Mol. Cell. Biol. *18*, 5404–5413.

Siebert, M., and Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. Nucleic Acids Res. *44*, 6055–6069.

Sieh, M., Bolen, J.B., and Weiss, A. (1993). CD45 specifically modulates binding of Lck to a phosphopeptide encompassing the negative regulatory tyrosine of Lck. EMBO J. *12*, 315–321.

Simmonds, P., Xia, W., Baillie, J.K., and McKinnon, K. (2013). Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla--selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics *14*, 610.

Simon, A. (2010). FastQC: A quality control tool for high throughput sequence data.

Singh, P., Lee, D.-H., and Szabó, P.E. (2012). More than insulator: multiple roles of CTCF at the H19-Igf2 imprinted domain. Front. Genet. *3*, 214.

Smith, M.L., Lopez, M.F., Archer, K.J., Wolen, A.R., Becker, H.C., and Miles, M.F. (2016). Time-Course Analysis of Brain Regional Expression Network Responses to Chronic Intermittent Ethanol and Withdrawal: Implications for Mechanisms Underlying Excessive Ethanol Consumption. PLOS ONE *11*, e0146257.

Stadler, M.B., Shomron, N., Yeo, G.W., Schneider, A., Xiao, X., and Burge, C.B. (2006). Inference of Splicing Regulatory Activities by Sequence Neighborhood Analysis. PLOS Genet. *2*, e191.

Staffa, A., and Cochrane, A. (1995). Identification of positive and negative splicing regulatory elements within the terminal tat-rev exon of human immunodeficiency virus type 1. Mol. Cell. Biol. *15*, 4597–4605.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., The RGASP Consortium, Hubbard, T.J., Guigó, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods *10*, 1177–1184.

Stoltzfus, C.M. (2009). Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. Adv. Virus Res. *74*, 1–40.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23.

Straube, J., Gorse, A.-D., Huang, B.E., and Lê Cao, K.-A. (2015). A Linear Mixed Model Spline Framework for Analysing Time Course 'Omics' Data. PLoS ONE *10*.

Stubbington, M.J., Mahata, B., Svensson, V., Deonarine, A., Nissen, J.K., Betz, A.G., and Teichmann, S.A. (2015). An atlas of mouse CD4+ T cell transcriptomes. Biol. Direct *10*.

Stutz, F., Bachi, A., Doerks, T., Braun, I.C., Séraphin, B., Wilm, M., Bork, P., and Izaurralde, E. (2000). REF, an evolutionary conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export. RNA N. Y. N *6*, 638–650.

Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D., et al. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput. Biol. *2*, e4.

Sun, H., and Chasin, L.A. (2000). Multiple Splicing Defects in an Intronic False Exon. Mol. Cell. Biol. *20*, 6414–6425.

Takahama, Y. (2006). Journey through the thymus: stromal guides for T-cell development and selection. Nat. Rev. Immunol. *6*, 127–135.

Takata, M.A., Gonçalves-Carneiro, D., Zang, T.M., Soll, S.J., York, A., Blanco-Melo, D., and Bieniasz, P.D. (2017). CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature *550*, 124–127.

Takata, M.A., Soll, S.J., Emery, A., Blanco-Melo, D., Swanstrom, R., and Bieniasz, P.D. (2018). Global synonymous mutagenesis identifies cis-acting RNA elements that regulate HIV-1 splicing and replication. PLOS Pathog. *14*, e1006824.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., and Burge, C.B. (2016). RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol. Cell *64*, 294–306.

Tang, M.W., and Shafer, R.W. (2012). HIV-1 Antiretroviral Resistance. Drugs *72*, e1–e25.

Tao, X., Constant, S., Jorritsma, P., and Bottomly, K. (1997). Strength of TCR signal determines the costimulatory requirements for Th1 and Th2 CD4+ T cell differentiation. J. Immunol. Baltim. Md 1950 *159*, 5956–5963.

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768.

Tartour, K., Appourchaux, R., Gaillard, J., Nguyen, X.-N., Durand, S., Turpin, J., Beaumont, E., Roch, E., Berger, G., Mahieux, R., et al. (2014). IFITM proteins are incorporated onto HIV-1 virion particles and negatively imprint their infectivity. Retrovirology *11*, 103.

Taylor, S.J., and Shalloway, D. (1994). An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis. Nature *368*, 867–871.

The FANTOM Consortium, Suzuki, H., Forrest, A.R.R., van Nimwegen, E., Daub, C.O., Balwierz, P.J., Irvine, K.M., Lassmann, T., Ravasi, T., Hasegawa, Y., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat. Genet. *41*, 553–562.

Tocchini-Valentini, G.D., Fruscoloni, P., and Tocchini-Valentini, G.P. (2011). Evolution of introns in the archaeal world. Proc. Natl. Acad. Sci. *108*, 4782–4787.

Tong, A., Nguyen, J., and Lynch, K.W. (2005). Differential expression of CD45 isoforms is controlled by the combined activity of basal and inducible splicing-regulatory elements in each of the variable exons. J. Biol. Chem. *280*, 38297–38304.

Topp, J.D., Jackson, J., Melton, A.A., and Lynch, K.W. (2008). A cell-based screen for splicing regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable exon 4. RNA N. Y. N *14*, 2038–2049.

Tress, M.L., Abascal, F., and Valencia, A. (2017a). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem. Sci. *42*, 98–110.

Tress, M.L., Abascal, F., and Valencia, A. (2017b). Most Alternative Isoforms Are Not Functionally Important. Trends Biochem. Sci. *42*, 408–410.

Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. *19*, 40.

Trinchieri, G., and Sher, A. (2007). Cooperation of Toll-like receptor signals in innate immune defence. Nat. Rev. Immunol. *7*, 179–190.

Trinchieri, G., Pflanz, S., and Kastelein, R.A. (2003). The IL-12 family of heterodimeric cytokines: new players in the regulation of T cell responses. Immunity *19*, 641–644.

Tulloch, F., Atkinson, N.J., Evans, D.J., Ryan, M.D., and Simmonds, P. (2014). RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. ELife *3*, e04531.

Turvey, S.E., and Broide, D.H. (2010). Innate immunity. J. Allergy Clin. Immunol. *125*, S24–S32.

UNAIDS (2014). The Gap Report.

Usami, Y., Popov, S., Popova, E., Inoue, M., Weissenhorn, W., and G Göttlinger, H. (2009). The ESCRT pathway and HIV-1 budding. Biochem. Soc. Trans. *37*, 181–184.

Usami, Y., Wu, Y., and Göttlinger, H.G. (2015). SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. Nature *526*, 218–223.

Usui, T., Nishikomori, R., Kitani, A., and Strober, W. (2003). GATA-3 suppresses Th1 development by downregulation of Stat4 and not through effects on IL-12Rbeta2 chain or T-bet. Immunity *18*, 415–428.

Vallejo, A.N., Brandes, J.C., Weyand, C.M., and Goronzy, J.J. (1999). Modulation of CD28 Expression: Distinct Regulatory Pathways During Activation and Replicative Senescence. J. Immunol. *162*, 6572–6579.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R.L., Mitchell, R., Johnson, M.C., Stephens, E.B., and Guatelli, J. (2008). The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein. Cell Host Microbe *3*, 245–252.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. ELife *5*, e11752.

Vernet, C., and Artzt, K. (1997). STAR, a gene family involved in signal transduction and activation of RNA. Trends Genet. TIG *13*, 479–484.

Voelker, R.B., and Berglund, J.A. (2007). A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. Genome Res. *17*, 1023–1033.

Wang, Z., and Burge, C. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, H.-Y., Xu, X., Ding, J.-H., Bermingham, J.R., and Fu, X.-D. (2001). SC35 Plays a Role in T Cell Development and Alternative Splicing of CD45. Mol. Cell *7*, 331–342.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics *23*, 1274–1281.

Wang, M., Zhao, Y., and Zhang, B. (2019). SuperExactTest: Exact Test and Visualization of Multi-Set Intersections. R package version 1.0.7.

Wang, W., Qin, Z., Feng, Z., Wang, X., and Zhang, X. (2013). Identifying differentially spliced genes from two groups of RNA-seq samples. Gene *518*, 164–170.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Swanstrom, R., Burch, C.L., and Weeks, K.M. (2009). Architecture and Secondary Structure of an Entire HIV-1 RNA Genome. Nature *460*, 711–716.

Weatheritt, R.J., Sterne-Weiler, T., and Blencowe, B.J. (2016). The ribosome-engaged landscape of alternative splicing. Nat. Struct. Mol. Biol. *23*, 1117–1123.

Wen, J., Chen, Z., and Cai, X. (2013). A Biophysical Model for Identifying Splicing Regulatory Elements and Their Interactions. PLoS ONE *8*.

Wensing, A.M., Calvez, V., Günthard, H.F., Johnson, V.A., Paredes, R., Pillay, D., Shafer, R.W., and Richman, D.D. (2014). 2014 Update of the drug resistance mutations in HIV-1. Top. Antivir. Med. *22*, 642–650.

Whisenant, T.C., Peralta, E.R., Aarreberg, L.D., Gao, N.J., Head, S.R., Ordoukhanian, P., Williamson, J.R., and Salomon, D.R. (2015). The Activation-Induced Assembly of an RNA/Protein Interactome Centered on the Splicing Factor U2AF2 Regulates Gene Expression in Human CD4 T Cells. PloS One *10*, e0144409.

Wickham, H. (2012). ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York).

Will, C.L., and Lührmann, R. (2011). Spliceosome Structure and Function. Cold Spring Harb. Perspect. Biol. *3*.

World Health Organization; UNAIDS; UNICEF. (20111). Global HIV/AIDS response: epidemic update and health sector progress towards universal access: progress report 2011.

Wu, Z., Jia, X., de la Cruz, L., Su, X.-C., Marzolf, B., Troisch, P., Zak, D., Hamilton, A., Whittle, B., Yu, D., et al. (2008). Memory T cell RNA rearrangement programmed by heterogeneous nuclear ribonucleoprotein hnRNPLL. Immunity *29*, 863–875.

Xin Wang, Kejun Wang, Guohua Wang, Sanford, J.R., and Yunlong Liu (2008). Model-based prediction of cis-acting RNA elements regulating tissue-specific alternative splicing. In 2008 8th IEEE International Conference on BioInformatics and BioEngineering, pp. 1–6.

Xu, Z., and Weiss, A. (2002). Negative regulation of CD45 by differential homodimerization of the alternatively spliced isoforms. Nat. Immunol. *3*, 764–771.

Yabas, M., Elliott, H., and Hoyne, G.F. (2015). The Role of Alternative Splicing in the Control of Immune Homeostasis and Cellular Differentiation. Int. J. Mol. Sci. *17*.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817.

Yap, K., and Makeyev, E.V. (2013). Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms. Mol. Cell. Neurosci. *56*, 420–428.

Yap, K., Lim, Z.Q., Khandelia, P., Friedman, B., and Makeyev, E.V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev. *26*, 1209–1223.

Yeo, G.W., Van Nostrand, E.L., Nostrand, E.L.V., and Liang, T.Y. (2007). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLoS Genet. *3*, e85.

Yi, R., Bogerd, H.P., and Cullen, B.R. (2002). Recruitment of the Crm1 nuclear export factor is sufficient to induce cytoplasmic expression of incompletely spliced human immunodeficiency virus mRNAs. J. Virol. *76*, 2036–2042.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinforma. Oxf. Engl. *26*, 976–978.

Yu, H., Wang, J., Sheng, Q., Liu, Q., and Shyr, Y. (2019). beRBP: binding estimation for human RNA-binding proteins. Nucleic Acids Res. *47*, e26–e26.

Zeng, C., and Hamada, M. (2020). RNA-Seq Analysis Reveals Localization-Associated Alternative Splicing across 13 Cell Lines. Genes *11*.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, Article17.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. *29*, 607–614.

Zhang, J., Kuo, C.-C.J., and Chen, L. (2012). VERSE: A Varying Effect Regression for Splicing Elements Discovery. J. Comput. Biol. *19*, 855–865.

Zhang, S., Wei, J.S., Li, S.Q., Badgett, T.C., Song, Y.K., Agarwal, S., Coarfa, C., Tolman, C., Hurd, L., Liao, H., et al. (2016). MYCN controls an alternative RNA splicing program in high-risk metastatic neuroblastoma. Cancer Lett. *371*, 214–224.

Zhang, X.H.-F., Leslie, C.S., and Chasin, L.A. (2005). Dichotomous splicing signals in exon flanks. Genome Res. *15*, 768–779.

Zhou, L., Chong, M.M.W., and Littman, D.R. (2009). Plasticity of CD4+ T Cell Lineage Differentiation. Immunity *30*, 646–655.

Zhou, Z., Qiu, J., Wen, L., Zhou, Y., Plocinik, R.M., Li, H., Hu, Q., Ghosh, G., Adams, J.A., Rosenfeld, M.G., et al. (2012). The Akt-SRPK-SR Axis Constitutes a Major Pathway in Transducing EGF Signaling to Regulate Alternative Splicing in the Nucleus. Mol. Cell *47*, 422–433.

Zhu, J., Guo, L., Watson, C.J., Hu-Li, J., and Paul, W.E. (2001). Stat6 is necessary and sufficient for IL-4's role in Th2 differentiation and cell expansion. J. Immunol. Baltim. Md 1950 *166*, 7276–7281.

Zhu, J., Jankovic, D., Grinberg, A., Guo, L., and Paul, W.E. (2006). Gfi-1 plays an important role in IL-2-mediated Th2 cell expansion. Proc. Natl. Acad. Sci. U. S. A. *103*, 18214–18219.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. *67*, 301–320.

(2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338.